DOCUMENT RESUME

ED 318 749 TM 014 763

AUTHOR	Medina, Noe; Neill, D. Monty
TITLE	Fallout from the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools. Third Edition (Revised).
INSTITUTION	National Center for Fair and Open Testing (FairTest), Cambridge, MA.
PUB DATE	Mar 90
NOTE	80p.
AVAILABLE FROM	FairTest, 342 Broadway, Cambridge, MA 02139 ($8.95).
PUB TYPE	Viewpoints (120) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Academic Achievement; Administrators; Construct Validity; Content Validity; Educational Assessment; Educational Change; *Educational Quality; Elementary Secondary Education; Equal Education; Multiple Choice Tests; Public Schools; *Standardized Tests; State Departments of Education; Surveys; *Testing Problems; Test Use
IDENTIFIERS	*Performance Based Evaluation

ABSTRACT

Standardized tests often produce results that are inaccurate, inconsistent, and biased against minority, female, and low-income students. Such tests shift control and authority into the hands of the unregulated testing industry and can undermine school achievement by narrowing the curriculum, frustrating teachers, and driving students out of school. It is contended that, as a whole, standardized tests do not measure much. Current standardized multiple-choice tests are severely flawed, based on assumptions about human ability that cannot be proven, with inadequate content and construct validity. An agenda for test reform calls for the use of "authentic" or "performance-based" evaluation, using properly constructed, validated, and administered tests. Schools, test-takers, and independent researchers should have access to data needed to verify test publishers' claims about their tests. Both test developers and test users have the obligation to promote proper, reasonable, and limited use of standardized tests as one of a series of assessment mechanisms. Appendix A presents results of a survey of officials from all states and from 56 sample school districts in 38 states focusing on the use of standardized tests. Appendix B discusses "authentic evaluation." A 72-item annotated bibliography is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED318749

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. Monty Neill/
FairTest

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Fallout From the Testing Explosion:

How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools

by Noe Medina and D. Monty Neill

with the staff of the
National Center for Fair & Open Testing
(FairTest)

FairTest
National Center for Fair & Open Testing
342 Broadway
Cambridge, MA 02139
(617) 864-4810

Third Edition (Revised)
March, 1990

BEST COPY AVAILABLE

A National Center for Fair & Open Testing Report

Fallout From the Testing Explosion:

How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools

by Noe Medina and D. Monty Neill

with the staff of the National Center for Fair & Open Testing
(FairTest)

FairTest Staff
Mike Friedman
Noe Medina
Nancy Murray
Monty Neill
Bob Schaeffer
Cinthia Schuman
Jamie Souder
Sarah Stockwell
John Weiss

© National Center for Fair & Open Testing (Fairtest)
Revised Third Edition, March 1990
Second Edition, October 1988
First Edition, June 1988

Contents

Introduction	3
Test Use in U.S. Schools	5
Problems with Standardized Tests	8
Impact on the Public Schools	24
Reform Agenda	33
Notes	37
Appendix A: Survey Results	47
Table 1: State-Mandated Tests	53
Table 2: District-Mandated Tests	54
Table 3: Test Use Calculations	57
Appendix B: Authentic Evaluation	58
Annotated Bibliography	61

FALLOUT FROM THE TESTING EXPLOSION:
HOW 100 MILLION STANDARDIZED EXAMS UNDERMINE EQUITY AND EXCELLENCE IN AMERICA'S PUBLIC SCHOOLS

"Standardized testing seems to have become the coin of the educational realm. In recent years, it seems that the aims of education and the business of our schools are addressed not so much in terms of curriculum . . . as in terms of what gets tested."[1]
- George Madaus & Walt Haney

Standardized tests dominate the educational landscape in contemporary America. Based on a recent study, the National Center for Fair & Open Testing (FairTest) conservatively estimates that public schools in the United States administered at least 100 million standardized tests to their 39.8 million students during the 1986-87 school year — an average of more than two and one-half tests per student per year. An estimate of 200 million—an average of five per year—is probably not extreme.

Public schools administered over 100 million standardized tests during the 1986-87 school year, an average of more than two and one-half per student per year.

Standardized test results have become the major criteria for a wide range of school decisions. Test scores limit the programs students enter or dictate the ones in which they are placed. Some tests are used to decide who will be promoted and who will be retained in grade; others determine which students graduate from high school. Test results also are used to assess the quality of teachers, administrators, schools and whole school systems.

Test proponents, of course, applaud these trends. They see tests as "objective" mechanisms to inject accountability into public schools and thereby improve student achievement, staff competence and educational quality. They see standardized exams as essential elements of the "School Reform Movement."

In fact, experience with standardized test use paints quite a different picture. Rather than being "objective" instruments, standardized tests often produce results that are *inaccurate, inconsistent, and biased* against minority, female and low-income students. Rather than promoting accountability, tests *shift control and authority* into the hands of an unregulated testing industry. By narrowing the curriculum, frustrating teachers, and driving students out of school, tests *undermine school improvement* rather than advance its cause.

Standardized tests often produce results that are inaccurate, inconsistent, and biased against minority, female and low-income students.

FairTest concurs with the National Academy of Education that "the nation has a right to know what students achieve, what schools

Relying on standardized tests will lead to a weaker, not stronger, educational system.

are doing, and what more should be done."[2] Standardized tests, even when properly constructed, validated, administered and used, can only play a limited role in this effort. Too often, however, standardized tests do not meet these basic standards. Moreover, the essential nature of these types of instruments makes them inadequate tools for assessing what is most valuable in education. As a result, relying on standardized tests as the primary criterion for making various school decisions will lead to a *worse, not better*, public understanding of the schools and a *weaker, not stronger*, educational system.

I. TEST USE IN U.S. SCHOOLS

During the 1986-87 school year, American educators reported that at least 93 to 105 million standardized tests or test batteries were administered to 39.8 million elementary and secondary public school students. This includes:

- 38.9 million standardized achievement, competency and basic skills tests administered to fulfill local testing mandates;
- 16.8 million standardized achievement, competency and basic skills tests administered in 42 states and the District of Columbia to fulfill state testing mandates;
- between 30 and 40 million standardized tests administered to compensatory and special education students;[3]
- between 1.5 and 1.75 million screening tests for kindergarten and pre-kindergarten students;[4] and
- between 6 and 7 million college and secondary school admissions, Graduate Equivalency Degree (GED) and National Assessment of Educational Progress (NAEP) tests.

This data was gathered by FairTest staff through a series of telephone interviews with officials from all 50 state departments of education, the District of Columbia school district and 56 sample school districts in 38 states. Additional information was gathered by examining recent surveys documenting the use of other standardized exams, including IQ tests, behavioral tests, readiness tests for young children and placement tests [see Appendix A].

FairTest interviewed education officials from all 50 states and the District of Columbia.

This estimate of 93-105 million tests is a conservative one. FairTest counted each test battery as one test. However, batteries usually include a number of separate exams: at least one each for math and reading, and frequently social studies, science and other subjects. If we had counted each exam within a battery as a separate test, the local- and state-mandated totals would easily double and the special education total would potentially double. The total also does not include tests administered to identify or place gifted or limited-English proficient students (for which there are no reliable figures). Nor does it include tests administered by private and parochial schools to their students.

Test use may also have been underreported by some school officials. For example, the Milwaukee Public Schools reported 64,500 tests administered in 1986-87. However, a detailed investigation conducted by the Milwaukee Assessment Task Force revealed a total of 484,956 tests administered in 1988-89 (each test in a battery was counted). Thus, rather than being tested at a rate of .67 tests per student per year, the rate is 5 per student per year. Milwaukee students in special programs were found to be tested far more heavily than students in regular classes.[5]

Given factors such as these, total standardized testing across the U.S. could exceed 200 million per year. Thus, in 12 years of schooling (plus kindergarten), a child could take, on average, some 60 standardized tests, or five per year.

The number of states which mandate testing has greatly increased in recent years.

In addition, the FairTest survey revealed that the number of states which mandate testing has greatly increased in recent years. Compared to the findings of a 50-state survey conducted by *Education Week* in 1985:

- the number of states requiring students to pass a standardized test for high school graduation increased from 15 in 1985 to 24 in 1987;

- the number of states employing standardized tests to determine whether students should be promoted to the next grade increased from 8 in 1985 to 12 in 1987; and

- the number of states using standardized tests as part of a state assessment program increased from 37 in 1985 to 42 in 1987.[6]

Southern states test most often.

The survey also revealed three significant patterns of standardized test use in public schools. First, ten Southern states (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina and Tennessee) administered standardized tests to fulfill state testing mandates at a rate more than twice that of schools in the remainder of the nation. In fact, the four states mandating the most tests per pupil (Kentucky, North Carolina, Alabama, and Georgia) were all located in the South.[7]

Second, seven states (Iowa, Minnesota, Montana, North Dakota, Ohio, Vermont, and Wyoming) which did not have a state testing mandate all have relatively small minority enrollments in their public schools (ranging from a low of 1% in Vermont to a high of only 15% in Ohio, compared to a national average of about 25%). Alaska is the only state without a testing mandate which also has a

substantial minority student population (primarily Native Americans). However, Alaska planned to institute a state testing program in the 1988-89 school year.[8]

Schools in larger districts administer more locally-mandated tests.

Finally, schools in larger school districts are likely to administer standardized tests to fulfill local mandates at a higher rate than schools in smaller districts. The largest school districts (those with student enrollments exceeding 100,000) administered tests at a rate one and one-quarter times that of medium-sized districts (those with enrollments between 25,000 and 100,000) and one and one-half times that of the smallest districts (those with enrollments less than 25,000). Within these overall trends, the rate of test administration

did vary considerably among districts.

II. PROBLEMS WITH STANDARDIZED TESTS

"The importance of understanding what it is that tests can and cannot tell us is critical. Not all tests are accurate measures of the skills and knowledge they purport to measure and even the more accurate tests are at best approximations."[9]
- Congressional Budget Office

Flaws in construction, validation, administration and use undermine claims of objectivity and produce test results that are inaccurate, unreliable or biased.

Standardized tests are consistently sold as scientifically developed instruments which simply, objectively and reliably measure student achievement, abilities or skills. In reality, there are serious problems in the construction, validation, administration and use of standardized tests and their results.

Standardized tests are constructed in ways that often guarantee biased results against minorities, females and low-income students. Test results are evaluated and scored in ways that are often at odds with modern theories of intelligence and child development. The test validation process is often inadequate and far from objective. Many tests are administered in an environment that undermines any claims they may have to being "standardized." Even those that adhere to "standard" administration practices may be biasing the results against minorities, low-income students and females by using examiners who are unfamiliar to the test-takers and by using tightly timed tests.

These flaws undermine testmakers' claims of objectivity and produce test results that are inaccurate, unreliable or biased. Ultimately, many tests fail to effectively measure test-takers' achievement, abilities or skills.

Test Bias

Standardized tests do not just reflect the impact of biases, they compound them.

Because most standardized tests are written by and for the middle- to upper-class white population, their results often fail to accurately measure the performance of those who do not fit this category. Testmakers claim that the lower test scores of racial and ethnic minorities and of low-income students simply reflect the bias and inequity that exists in American schools and society. While such biases and inequities certainly exist, standardized tests do not just reflect their impact, they compound them.

Joseph Gannon provided documentation for this conclusion in a 1980 study for the National Conference of Black Lawyers. Gannon

college seniors from the same universities and with comparable undergraduate grade point averages. Even after controlling for these characteristics a gap of 100 points remained when black and Hispanic scores were compared with those of white students even though they had demonstrated equal academic ability in college.[10]

Researchers have identified several characteristics of standardized tests which could bias results against minority and low income students. Each characteristic reflects the middle- to upper-class white focus of such tests. As such, test results are as much a measure of race/ethnicity or income as they are of achievement, ability or skill.

Researchers have identified several characteristics of standardized tests which could bias results against minorities and low income students.

Some of these characteristics also lead to gender bias in standardized tests, which affects both males and females. Among very young children, some tests appear to be biased against boys. On the other hand, among older children and adolescents, most bias affects girls.[11]

Several of the characteristics that bias test results relate to language. To communicate their level of achievement, ability or skills, test-takers must understand the language of the test. Obviously, tests written in English cannot effectively assess the achievement, skills or abilities of students who primarily speak Spanish or some other language.

Many groups of English-speakers are affected by a similar, but more subtle, form of language bias. Most standardized tests are written in an elaborated, stylized language rather than the simple and common vernacular. Researchers have discovered that the use of such forms of English prevents tests from accurately measuring achievement, ability or skills of students who use nonstandard English dialects. This includes speakers of African-American, Hispanic, white southern, Appalachian, and working-class dialects.[12]

A related type of bias stems from stylistic or interpretive differences related to culture, income or gender. For instance, the word "environment" is often associated by black students with terms such as "home" or "people." White students tend to associate the word with "air," "clean" or "earth." Neither usage is wrong; one simply centers on the social environment while the other centers on the natural one. Unfortunately, on a standardized test only one of these two usages, generally the one reflecting the white perspective, will be acceptable.[13]

On a standardized test, only one usage, generally the one reflecting the white perspective, will be acceptable.

Similarly, researchers have discovered that individuals exhibit "different ways of knowing and problem-solving" which reflect different styles, not different abilities. These differences are often

correlated with race/ethnicity, income level and gender. Yet standardized tests assume that all individuals perceive information and solve problems in the same way.[14]

Standardized tests assume that all individuals perceive information and solve problems in the same way.

For example, middle class whites are more apt to be trained, simply through cultural immersion, to respond to questions removed from any specific context and to repeat information the test-taker knows the questioner already possesses. Heath found that working class black children, in their communities, were rarely asked the sorts of questions in which the questioner already knew the answer, the sort found on tests.[15] Again, assumptions about the universal application of one particular style limits the validity of test results.

Another cause of test bias is apparent in the content of questions which assume a cultural experience and perspective which not all children share. "Correct" answers to such questions usually reflect the experiences, perspectives and knowledge of children and adults from a white middle- to upper-class background. Answers which draw upon the different experiences, perspectives and knowledge of racial/ethnic minorities, children from poor families, and children from inner city or rural backgrounds are ignored. Although some answers may be correct in these different geographical or cultural contexts, they are generally counted "incorrect" on the test.

The WISC-R IQ test, for example, asks "What is the thing to do when you cut your finger?" The best response, according to the test, is to "put a bandaid on it." Partial credit is also given for a response of "go to the doctor (hospital)." No credit is awarded for responses of "cry," "bleed" or "suck on it." Minority children usually perform poorly on this item. A few years ago a Baltimore, Maryland sociologist asked several inner-city youths why they answered the question the way they did. She found that many of these kids answered "go to the hospital" because they thought that "cut" meant a big cut. When the children were told that "cut" meant "little cut," almost all then responded, "Put a bandaid on it."[16]

Research on Mexican Americans, African Americans and girls all reveal that "items with content reference of special interest" to each group seem to improve their test scores.

Finally, students tend to perform better on tests when they identify with the subjects of the test questions. Research on Mexican-Americans, African-Americans, and girls all reveal that "items with content reference of special interest" to each group seem to improve their test scores.[17] Unfortunately, standardized tests remain dominated by questions about and for white middle- to upper-class males.[18]

Nonetheless, test companies maintain that they effectively screen out biased questions. Though they subject items to review by experts who can supposedly detect bias, such screening is of low accuracy, and many biased items cannot be detected in this way.[19] Though most major test-makers also apply some form of statistical bias-detection procedure, even when bias is found items are not necessarily removed. Moreover, the procedures themselves are problematic.

Differential Item Functioning does not detect bias in the test as a whole.

Bias-reduction techniques, such as Differential Item Functioning, (DIF), generally attempt to match test-takers by "ability" and then locate those items that test-takers who are of similar "ability" but differ by race, ethnicity, or gender answer correctly at different rates. The items which exhibit large race, ethnicity or gender differentials are then flagged for further review, but not necessarily discarded from the test even if the differential answer rate is very large. However, the procedure used to match test-takers by "ability" is to match them by their overall scores on the test. Since the overall test score is nothing more than the sum of all test items, this procedure is obviously circular (as the Educational Testing Service has admitted) and inadequate for detecting any systematic bias in the test as a whole. The fundamental problem is that the procedure assumes the very thing it ought to be checking for.[20]

Given the role of knowledge and language in creating culture, there is no reason to believe a totally "culture-free" test can be constructed. Tests however, can be better designed so as to reduce their discriminatory impact by careful selection of content that is less likely to be unfamiliar to minorities or by use of bias reduction techniques such as the "Golden Rule" procedures.[21]

All "objective" really means is that the test can be scored without human subjectivity, by machines.

It is important to note that the use of standardized tests is often defended on the grounds of "objectivity." But all "objective" really means is that the test can be scored without human subjectivity, by machines. As Banesh Hoffman noted in *The Tyranny of Testing*, "the term 'objective test' is a misnomer. The objectivity resides not in the test as a whole but merely in the fact that no subjective element enters the *process* of grading once the key is decided upon."[22]

Bias can still creep into the questions themselves. What content and which items to include on the test, the wording and content of the items, and the determination of the correct answer, as well as how the test is administered and the uses made of test scores, are all decisions made subjectively by human beings. In fact, the purported objectivity of tests is often no more than the standardization of bias.

The replacement of the potential bias of individual judgement with the numerical bias of an "objective" test is not progress.

objectivity of tests is often no more than the standardization of bias. The replacement of the potential bias of individual judgement with the numerical bias of an "objective" test is not progress.

Test Construction

The ability of standardized tests to accurately report students' knowledge, abilities, or skills is limited by assumptions that these attributes can be isolated, sorted to fit on a linear scale, and reported in the form of a single score. Gould labels these the fallacies of *reification* (treating "intelligence" as though it were a separable unitary thing underlying the complexity of human mental activity) and *ranking* ("our propensity for ordering complex variation as a gradual ascending scale" using a number). He concludes, "Thus, the common style embodying both fallacies of thought has been quantification, or the measurement of intelligence as a single number for each person."[23] As Levidow remarks, the process works so that "Without anyone having to claim that IQ scores represent the quantity of a thing, it appears that way by virtue of assigning a number to each testee and then comparing those numbers through a distribution curve."[24]

Many of the assumptions and structures of achievement tests are based on IQ tests and operate in the same way. For example, assumptions regarding the linear sorting of students are common to both.[25] Such assumptions are at odds with contemporary research on child development, which emphasizes diversity in the nature and the pace of child development.[26] Child language research, for example, demonstrates that "some children develop the use of pronouns before the development of an extensive noun vocabulary. For others, the reverse pattern of development occurs." Neither is considered to reflect a learning disorder or disability. They simply reflect variations in development patterns.[27] Thus, standardized tests mislabel many, if not most, individuals who exhibit developmental patterns that differ from a defined "norm" (based on majority group practice) as being delayed or disordered in their development. Normal human variation, then, is defined as a problem.[28]

Standardized tests mislabel many individuals who exhibit developmental patterns that differ from a defined "norm."

The use of a linear scale not only creates false differences, it also can mask real differences. Assume, for example, that one student can compute using addition, subtraction, multiplication or division, but is unable to apply those concepts to fractions. Meanwhile, another student can compute with either whole numbers or fractions, but has difficulty with multiplication and division. If a mathematics test included four questions on whole numbers and four on fractions and each question required the student to employ a different type of

score (50%). Yet, their identical scores would mask real differences in skills.

The simple fact is that, while our knowledge of thinking, learning, teaching, and child development has grown over recent years, many standardized tests have not. Despite claims that testing is now more advanced and scientific, Oscar Buros noted that "little progress has been made in the past fifty years — in fact in some areas, we are not doing as well. Except for the tremendous advances in electronic scoring, analysis and reporting of test results, we don't have a great deal to show for fifty years' work. Essentially, achievement tests are being constructed today in the same manner they were fifty years ago. . ."[29]

While our knowledge of thinking, learning, teaching, and child development has grown over recent years, many standardized tests have not.

The same can be said for I.Q. and school readiness testing. The WISC-R "has remained virtually unchanged since its inception in 1949. . . Developments in the fields of cognitive psychology and neuroscience have revolutionized our thinking about thinking, but the WISC-R remains the same."[30]

Resnick and Resnick point out that behaviorist and associationist psychological theories from early in the twentieth century underlie standardized tests.[31] These theories assume that knowledge can be decomposed or broken into separable bits, taken out of any context, and learned as an accumulation of these bits. The bits can then be tested one by one and without context. This is sometimes called the "banking" or "empty vessel" theory of learning. While few psychologists hold to these theories today, these assumptions remain built into standardized testing.

Test constructors not only erroneously presume that the knowledge, skill or ability being measured is one-dimensional and decomposable, but also that it tends to be distributed according to the statistical "normal" bell-shaped curve. The bell-shaped curve is used for statistical convenience, not because any form of knowledge or ability is actually distributed in this manner.[32]

Again, modern theories emphasize the complexity of human intelligence. Researchers have observed that knowledge, learning and thinking have multiple facets, and that a high level of development in one area does not necessarily indicate a high level of development in others.[33] Unitary test scores and linear scaling of scores ignore the true complexity and thus provide a deceptive picture of individual achievement, ability or skills. This is a fundamental problem underlying all standardized tests in education.

Standardized tests ignore the true complexity of human intelligence.

Creation of a norm-referenced curve exaggerates the differences among people.

In order to construct a "normal curve," test-makers must separate test-takers along a continuum. Thus, a norm-referenced test must not have many questions all test-takers can answer and must have some that almost none can answer. Creation of the curve exaggerates the differences among people. In general, test-makers discard those questions on which low-scoring test-takers do well but high-scorers do poorly.[34] As a result, an item on which African-Americans do particularly well but whites do not is likely to be discarded for the compound reason that African-Americans are a minority group in the U.S. and generally score low. Even if minorities are included in the test companies' samples consistent with their portion of the overall population, at least three quarters of the sample is white. Moreover, minorities are disproportionately among the low-scoring group. Thus, questions that might favor minorities are apt to be excluded for not fitting the "required" statistical properties of the test.

For standardized achievement tests, problems also arise when test publishers use national test score averages ("norms") as reference points for interpreting student performance. Using these norms, schools can determine, for example, that a certain test score ostensibly represents performance at the 65th national percentile, i.e. higher than 65% of all other students.

Test norms are developed by administering the test to a group of students which, in theory, represents the national student population. However, some widely used tests have been normed on small, unrepresentative populations. The popular Gesell Preschool Test, for example, employed a normative sample "composed of 40 girls and 40 boys at each 6-month age level from 2 through 6 years, for a total sample of 320 girls and 320 boys . . .however, 'nearly all were Caucasians and all resided in the state of Connecticut.'"[35]

Some widely used tests have been normed on small, unrepresentative populations.

The WISC-R was normed using a population of 100 boys and 100 girls for each of its age levels from 6-1/2 to 16-1/2. According to the test publisher, the norming group was representative of the national population as of 1970 with respect to race, geographic region, occupation of head of household, and urban-rural residence.[36] Aside from the fact that the national population has changed considerably since 1970, this means that the test is normed for African-Americans using only 10 boys and 10 girls at each grade level and for Hispanics using only 5 or 6 each of boys and girls.

Norms also may be distorted even if samples of a thousand or more students for each level are used (as is the case with major

standardized achievement batteries) because school systems volunteer to participate in the norming process. The quality of the norming sample therefore depends on how accurately the self-selected schools reflect national performance. Test publishers typically fail to report the extent of this potential variance.[37]

A recent study by Friends for Education raises additional doubts about the validity of the national norms. That organization "discovered that no state is below average at elementary level on any of the six major nationally normed, commercially available tests" and blamed inaccurate initial norms and teaching to the test for the inflated scores.[38]

"No state is below average at elementary level on any of the six major nationally normed, commercially available tests."
—Friends for Education

Critics have noted that the report released by Friends for Education contains inaccuracies and that many of the tests' statistical procedures were interpreted incorrectly. However, these flaws do not undermine the major finding of the report (dubbed the Lake Wobegon Phenomenon after Garrison Keillor's mythical community in which "all the children are above average"). In a recent analysis, Daniel Koretz reaffirmed the report's finding and conclusion that outdated norms are in part to blame for overstating student performance.[39] A subsequent study conducted for the Center for Research on Evaluation, Standards and Student Testing (CRESST) concluded that the "Lake Wobegon effect" does exist and argued that "teaching to the test" is a primary cause.[40]

Test Reliability

Publishers' claims that standardized tests exhibit a high level of reliability do not necessarily mean that test results will be similar in successive administrations. Rather, for the testing industry, "reliability" is a technical term encompassing several different concepts.

Publishers' claims that standardized tests exhibit a high level of reliability do not necessarily mean that test results will be similar in successive administrations.

The term *can* mean consistency of test scores over time, the popular perception of reliability. However, it can also refer to the consistency between scores on the entire test and responses to individual test questions, between two halves of the test, or between different forms of the same test. Despite sharing the same term, the concepts are certainly not interchangeable, even though each measures some aspects of the "range of fluctuation likely to occur in a single individual's score as a result of irrelevant, chance factors."[41]

Psychometricians have developed a variety of statistical approaches to measure these different types of reliability. These approaches all report their results as "reliability coefficients" (numbers generally ranging from 0 to 1). For most standardized tests, these re-

generally ranging from 0 to 1). For most standardized tests, these reliability coefficients are very high — often exceeding .8 and .9.[42] Test companies often report internal or inter-form reliability rather than consistency over time (which usually has a lower reliability), even where the latter may be more important than the former.[43]

Nevertheless, significant scoring differences can still result. Anne Anastasi presents an example in her text, *Psychological Testing*, of an IQ test with a reliability coefficient of .89 and a standard deviation of 15. A student administered that test twice would be likely to score up to 13 points higher or lower on a retest. (There is also a chance a student could score even higher or lower). Thus a school system could deny entry into a "Gifted & Talented" Program requiring a score of 130 on this test to a student scoring 117 when the same student could, within the bounds of reliability, readily score 130 on a readministration of the same test.[44]

The publisher of the Gesell does not report any reliability data for its tests.

Similarly, Kaufman and Kaufman found that the reliability of the Gesell School Readiness Test meant that "a child measured to be four and one-half years old developmentally and unready for school could very likely be five and fully ready."[45] The publisher of the Gesell, in fact, does not report any reliability data for its tests.[46]

The effect of imperfect reliability on decision making is expressed in various ways. The "error of measurement" means that decisions made using test scores with imperfect reliability will include both false positives (those who will be included but who should not be) and false negatives (those who will be excluded but should have been included). For example, arbitrary use of test scores will assign some children to a remedial program who do not belong there, while excluding others who should be in the program. Setting the "cut score," the point at which to include or exclude, high will lead to more false negatives, while setting it low will lead to more false positives. The setting of the cut score is an ultimately subjective act by test makers or test users.

Arbitrary use of test scores will assign some children to a remedial program who do not belong there, while excluding others who should be in the program.

Test-makers also can compute a test's "margin of error" and its "standard error of difference." The Scholastic Aptitude Test (SAT), for example, has a "standard error of measurement" of 67 points on the combined test (which has a scale that runs from 400 to 1600). The error of difference is 144 points. That is, two test-takers' scores must differ by at least that much before it can be said their abilities (as measured by the test) differ.[47] Decisions on college entrance, however, are sometimes based on SAT score differences as small as 10 points.

Even with tests which report overall high reliability, the reliability of test sub-sections may be much lower. On many tests, reliability is also lower for children below the third grade.[48] Thus the chance for error increases when such tests are used for placing young children or when decisions are made based on sub-test scores. Given the reality of test reliability, it is not surprising that the Congressional Budget Office noted, "one indication of the limitations of standardized results is often marked disparities in the results they yield."[49]

On many tests, reliability is lower for children below the third grade.

Analysis of test reliability leads to one inescapable conclusion: *No test has reliability high enough that it can be used as the sole or primary basis for making decisions about students.*[50]

Test Administration

Educators, researchers and members of the public generally assume that standardized tests are administered in a standardized context under relatively uniform conditions. Anastasi emphasizes the importance of such a controlled setting: "Even apparently minor aspects of the testing situation may appreciably alter performance. . . In general, children are more susceptible to examiner and situational influences than are adults; in the examination of preschool children, the role of the examiner is especially crucial."[51]

Research has demonstrated that tests are administered in far from "standard" conditions.

In fact, recent research has demonstrated that tests are administered in far from "standard" conditions. One study concluded that "the actual context [of test administration] often includes confusion, anxiety, behavioral resistance, negative attitudes toward testing on the part of staff and students, lack of properly trained test examiners, developmentally or educationally immature children, and other institutional problems that are endemic to many schools."[52] Because reliability coefficients *do not* take into account the possibility that tests are administered under such variable situations, the reliability of test scores, particularly for tests administered to very young children, is likely to be lower than estimates computed from controlled studies and reported by test publishers. Moreover, many of the ideal "standard" conditions called for by test developers may actually place certain groups of students at a disadvantage. For example, the use of unfamiliar test examiners reduces test scores of low socio-economic status (SES) and black students, but does not affect the scores of high SES students. This factor alone can account for half of the difference in I.Q. test scores between low and high SES students.[53] Similarly, the time limitations associated with most standardized tests harm minorities and women. Generally, these groups appear to cope less effectively with the pressures inherent in a time-limited test than do white males.[54]

The use of unfamiliar test examiners reduces test scores of low SES and Black students, but does not affect the scores of high SES students.

Test Validity

The validity of any standardized test depends entirely on its context.

"A test," write Airasian and Madaus, "is a sample of behaviors from a domain about which a user wishes to make inferences.... Test validity involves an evaluation of the correctness of the inferences about the larger domain of interest."[55] Thus, the validity of a standardized test tells us whether the test measures what it claims to measure, how well it measures it, and what can be inferred from that measurement. Test validity cannot be measured in the abstract. It can only be determined in the context of the specific uses to which a test's results will be put. Thus, information and conclusions regarding a test's validity in one context may not be relevant and applicable in different contexts.

Like reliability, the term "validity" encompasses several concepts:

•*Content-related validity* determines whether the test questions relate to the trait or traits the test purports to measure.

•*Criterion-related validity* compares test performance (for example, on a reading test) against a standard that independently measures the trait (such as reading ability) the test purports to measure. Criterion validity takes two forms, concurrent and predictive. Concurrent validity describes how well test results coincide with another measure of the same trait made at the same time. Predictive validity examines how well the test forecasts performance on another measure of the tested ability when assessed at a later date.

•*Construct-related validity* examines how well a test actually correlates with the underlying theoretical characteristics of the trait it purports to measure. For example, does the test accurately measure "academic ability" or "competence." This form of validity is rarely reported by test makers even though it is essential for assessing how useful and accurate a test will be in practice.[56]

The types of evidence required to demonstrate a test's validity will differ depending on how test results are to be used.

The types of evidence required to demonstrate a test's validity will differ depending on how test results are to be used. For example, an achievement test used to determine how well students have learned math would require content-related validity. If used to predict students' future math performance, the same test would also require criterion-related predictive validity. A test should only be used for purposes for which it has been adequately validated. Too often, test are used invalidly.

To be content valid at the level of sophistication of the appropriate domain, the test must adequately include what the domain covers. This is so difficult to do within the multiple-choice format that it is, essentially, not done (see discussion below, pp. 21-22).

Lack of adequate content validity can have wide-ranging effects. For example, if a U.S. history test only measures factual recall and the test is used to guide curriculum (as is increasingly the case), then not only will most of the real content of history not be measured, it will be excised from the curriculum.

Lack of adequate content validity can have wide-ranging effects.

Test developers (both commercial institutions and governmental agencies) generally validate a test's use by asking subject area experts to make qualitative judgments about the relationship between individual test items and the trait(s) the test seeks to measure. The selection of test items typically is done by panels of experts who review textbooks for content, draft items and then review them—a method occasionally referred to as BOGSAT (Bunch Of Guys Sitting Around a Table). Essentially, the subjective views of individuals are aggregated to design a test whose content is labelled "objective" and comprehensive.

Because each test item must be one that reasonably should be on the exam, experts are asked whether the item should be included. This is a simple, affirmative format. However, what content validity studies should examine are disconfirming hypotheses: What is not included? Is the overall balance of the items adequate to cover the content? Given the limited number of questions, is the content range a fair approximation of the domain? Such disconfirming questions generally are not asked during test development.[57]

Many test developers do not go beyond content-related validity.[58] For example, the widely-used and highly-respected Iowa Test of Basic Skills "is somewhat lacking when it moves beyond content validity into other validity realms."[59] Similar comments are made by reviewers of other standardized achievement tests.[60]

Unfortunately, many test developers do not go beyond content-related validity.

Test developers who do go beyond content-related validity generally rely upon other tests to demonstrate criterion-related and construct-related validity. For example, Mitchell demonstrated the predictive validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis by correlating scores on those tests with scores on the Stanford Achievement Test. However, she failed to explain what the Stanford Achievement Test measured and how validly it did so.[61]

Another approach to demonstrating criterion-related validity relies upon comparisons of test scores with teachers' grades. This, however, undermines a major selling point of standardized tests — that they are an objective substitute for overly subjective teacher judgments.[62]

The question, then, is whether the test is more reliable and valid than are teachers' judgments or some other plausible measure of ability or achievement.

If test scores and grades agree completely, then why have the tests at all? If they differ significantly, which is better and how do we know? The simple claim of test "objectivity" is insufficient. The question, then, is whether the test is more reliable and valid than are teachers' judgments or some other plausible measure of ability or achievement.[63] This last point is important because test-makers will argue that even with low validity, tests can improve decision-making as compared with pure chance. However, teacher judgments and other high-quality alternatives are *not* decisions equivalent to pure chance.

Validity, like reliability, can be measured by statistical methods which produce numbers called validity coefficients. For many standardized multiple-choice tests, validity coefficients are quite low, and even high coefficients can still result in significant margins of error. "Although various readiness tests are correlated with later school performance, predictive validities for all available tests are low enough that 30 to 50 percent or more of children said to be unready [for first grade] will be falsely identified."[64] The predictive validity of many developmental screening tests, exams that purport to indicate children's possible disabilities, is also low and often determined by comparing one test with another.[65]

"Predictive validities for all available tests are low enough that 30 to 50 percent or more of children said to be unready [for first grade] will be falsely identified."
—Shepard & Smith

Another major concern about predictive validity is whether the performance predicted is, in fact, created through a self-fulfilling prophecy. If a child scores low on a test and is then placed in a program in which he or she is not challenged and does not progress, is a low score on a subsequent achievement test a measure of the accuracy of the first test or only the result of placement in a slow track? Predictive validity studies do not address this question.

The constructs underlying tests are often entirely outdated, as is the case with IQ, readiness and achievement tests that are based on early twentieth-century psychology. Often, a test will purport to measure one thing when, in fact, it measures another. Deborah Meier, Principal of Central Park East Secondary School in Manhattan, argues that reading tests do not measure reading but rather measure "reading skills," such as phonics decoding, which is not the same thing.[66] That is, the tests are based on a faulty understanding of reading and learning to read. As Jerome Harste puts it, "Standard-

ized testing is hopelessly out of date given what we know about reading."[67]

What is true of reading is true across the board: the tests do not explain how people learn nor what they know beyond a very narrow and limited area. The tests also misidentify the content and the construct they purport to measure. As a result, the real knowledge and abilities of our students are not measured.[68]

Tests do not explain how people learn nor what they know beyond a very narrow and limited area.

This is true not only when standardized exams are used to test individuals, but also when used to assess programs. As Airasian and Madaus write, "Are traditional standardized achievement tests construct valid in terms of inferences about school or program effectiveness? In general, the answer is no."[69] The lack of construct validity has a direct impact on teaching when curriculum becomes dominated by testing.

In the work of leading psychometric theoreticians, construct validity has become the essential core of validity, subsuming content and criterion validity.[70] In large part, this is because questions about the meaning of the content or the effects of the prediction enter the realm of underlying hypotheses, theories and assumptions. Tests are not constructed and used independent of theories of knowledge, ability and performance, as well as theories about the domain to be measured. (For example, the domain of history must be conceptualized to provide a construct that can be measured.) The relationships among theories, tests and test use should be examined as part of construct validity studies. Typically, as indicated above, either the constructs are not considered at all or they are woefully inadequate or outdated.

Messick, among others, has argued that the construct validity of a test cannot be considered outside of social or educational values and the consequences of its use.[71] Though Messick himself maintains that we must distinguish between those adverse social consequences of test use that are attributable to the test and those that are not,[72] this expansion of the concept of construct validity opens the entire enterprise of testing to serious question. *If the general social results of testing are harmful, and the harm can be traced to the nature and structure of the tests, then testing should, by its own terms, be rejected as lacking in validity.*[73]

The construct validity of a test cannot be considered outside of social or educational values and the consequences of its use.

From this perspective, reliance on the multiple-choice format itself limits the construct validity of many tests. For example, this format appears fundamentally incapable of adequately measuring what are now commonly referred to as "higher order thinking

The multiple-choice format itself limits the construct validity of many tests. For example, this format appears fundamentally incapable of adequately measuring what are now commonly referred to as "higher order thinking skills.

skills."[74] In part, this is because the format requires that there be only one right answer. Some questions on a test may have, by accident, more than one correct answer. But more fundamentally, the multiple-choice format precludes the possibility of designing problems that have more than one correct response or more than one method of solution. Yet problems of these types are necessary to reflect the very essence of complex thinking. Further, multiple-choice questions are carefully constructed and defined, while real-world problems are usually messy and ill-structured. Nor are multiple-choice tests useful for explaining how students think as they solve problems.[75] Multiple-choice testing is not based on evaluating real work done under real conditions; that is, it is not based on students' authentic performances.

The theoretical justification for multiple-choice questions resides in behaviorist and associationist psychology.[76] A test constructed on behaviorist or similar theories may correlate well with behaviorist theory. However, because that theory is itself false, the test results can be of limited use at best and dangerous at worst because their use reinforces incorrect views of the content and of how people learn. Thus, resulting adverse social effects are due at least in part to the construct of the test and the test is invalid.

Finally, Johnston argues that the philosophy of science underlying the very concept of validity presumes a model of education in which the student and the teacher are both objects. This model, he charges, disempowers student and teacher, with detrimental effects to both as well as to education and society. What is needed, he concludes, is a different conception of science connected to a fundamentally different educational practice—different values, different consequences, and a different conception of what is valid.[77]

In general, the content, criterion and construct validity of standardized tests is so limited as to make most test use invalid.

In general, the content, criterion and construct validity of standardized tests is so limited as to make most test use invalid. To statisticians or test developers, the reliability and validity of their tests may seem adequate. However, the degree of error remains so high that any decision made about a child based on a single instrument risks harmful consequences. The content of the tests tends to be extremely narrow, due in large part to the multiple-choice format. The predictive validity of the tests often resides in circular logic or self-fulfilling prophecy. And test construct validity is generally absent or based on outmoded theories of learning and knowledge. Because of the enormous educational damage caused by reliance on standardized tests (discussed below), it should now be up to the test-makers to prove that any extensive use of their instruments is valid.

Summary

Even a cursory examination of current standardized tests reveals many flaws:

•Race, class, linguistic and gender bias may lead to incorrect scores and thus incorrect treatment of many groups and individuals.

•The assumptions about "intelligence" and how people learn are outdated and designed to fit statistical procedures, not reality.

•Norming is often inadequate or misleading and the norms are frequently out-of-date.

•The reliability of many tests is too low for accurate educational decision-making, particularly for young children.

•Variations in test administration can reduce test reliability, particularly for minorities and low-income children.

•The multiple-choice format is incapable of measuring higher order skills, thinking or creativity.

•Although proper test use requires different types of validity studies, test-makers generally perform only a limited content-related validity study. Validity theory itself now calls into question the validity of the entire testing enterprise.

Taken as a whole, standardized tests do not measure much, and what they do measure, they measure inadequately. Thus, tests should be used only with great caution. Unfortunately, in too many American schools that is simply not the case.

Taken as a whole, standardized tests do not measure much, and what they do measure, they measure inadequately.

III. THE IMPACT OF TESTING ON THE PUBLIC SCHOOLS

"The adverse impact of an educational system based on tests may be far worse than anticipated. A lesson is to be learned from an English experiment [in 1863] which utilized tests as a basis for accountability . . . What transpired was nothing short of disaster. . . Almost all courses that were not addressed in the test were dropped and reading materials were limited to those that appeared on the tests. . . Some teachers quickly left the system and prospective teachers were reluctant to enter a profession that operated in this manner. Needless to say 'payment by results' was abandoned in fairly short order."[78]

—Mary Dilworth

American public schools have begun to treat standardized tests as the all-purpose answer for promoting educational improvement and ensuring school accountability.

Traditionally, standardized tests have been one of several educational tools used to assess student achievement and to diagnose their academic strengths and weaknesses. In recent years, American public schools, like their British counterparts more than a century ago, have begun to treat standardized tests as the all-purpose answer for promoting educational improvement and ensuring school accountability. In the process, standardized tests have become the primary or sole criterion used by public schools for making a number of decisions affecting students, teachers and schools. In many schools, standardized tests serve as gatekeepers for:

- assignment to special education or remedial programs;
- admission to gifted and talented or accelerated programs;
- grade promotions;
- high school graduation;
- merit pay awards to teachers;
- teacher certification and recertification;
- allocation of funds to schools or school systems; and
- school system certification and decertification.

Reliance on standardized tests as educational gatekeepers is growing.

Reliance on standardized tests as educational gatekeepers is growing. In just two years, 1986-88, the number of states using tests to determine student promotions increased from 8 to 12. Similarly, the number of states using tests to determine eligibility for high school graduation increased from 15 to 24.

Given the limited range of skills and knowledge measured by standardized tests, the impact of race, ethnicity, income and gender on test results, and questions regarding their proper construction, validation and administration, the use of standardized tests as the primary or sole criteria for making any "high stakes" decision is reckless. Moreover, as standardized tests have become the all-powerful gatekeepers of American education, they have affected educational goals and curriculum, student progress and achievement and local control — and created a new set of problems in each area.

The use of standardized tests as the primary or sole criteria for making any "high stakes" decision is reckless.

Impact on Educational Goals and Curriculum

Children go to school not just to learn basic academic skills, but also to develop the personal, intellectual and social skills to become happy, productive members of a democratic society. Unfortunately, the current emphasis on standardized tests threatens to undermine this educational diversity by forcing schools and teachers to focus on quantifiable skills at the expense of less easily quantifiable academic and non-academic abilities.

This is particularly true for young children. As the National Association for the Education of Young Children (NAEYC) recently noted: "Many of the important skills that children need to acquire in early childhood — self-esteem, social competence, desire to learn, self-discipline — are not easily measured by standardized tests. As a result, social, emotional, moral, and physical development and learning are virtually ignored or given minor importance in schools with mandated testing programs."[79]

"Social, emotional, moral, and physical development and learning are virtually ignored in schools with mandated testing programs."
–NAEYC

Many schools have embarked on a single-minded quest for higher test scores even though this severely narrows their curriculum:[80]

•Deborah Meier noted that when synonyms and antonyms were dropped from the New York City test for word meaning, teachers promptly dropped academic material that stressed them. She also noted that students read "dozens of little paragraphs about which they then answer multiple-choice questions" — an approach that duplicates the form of the standardized tests the students take in the spring.[81]

•Gerald Bracey, former Director of Research, Evaluation and Testing in the Virginia Department of Education, noted that some teachers did not teach their students how to add and subtract fractions because the state's minimum competency test included ques-

tions on multiplication and division of fractions, but not on their addition and subtraction.[82]

• In one Georgia school, the goal in essay writing is to produce "a five-paragraph argumentative essay written under a time limit on a topic about which the author may or may not have knowledge, ideas, or personal opinions." Not surprisingly, this exercise exactly matches the requirement for the Georgia Regents' Test essay exam.[83]

Test preparation can begin months prior to the test.

Sometimes, the curriculum is narrowed simply because "testing takes time, and preparing students for testing takes even more time. And all this time is time taken away from real teaching."[84] In fact, test preparation can begin months prior to the test. Susan Harman of the American Reading Council has observed classrooms discarding all other curriculum in January to prepare for April testing in New York City.[85] The study by Smith, *et al. . .*, shows schools in Arizona focusing on test preparation in January for April testing.[86] A separate study for the Arizona Department of Education showed the results of the focus on the multiple-choice tests: three-fourths of the state curriculum remained untaught because it was not covered by the tests.[87] The essential reason appeared to be the pressure teachers felt to raise test scores.

"The curriculum falls in line with the test, and the test becomes the curriculum."
--P.S. Hlebowitsch

Unfortunately, a closer link between tests and curriculum has become a very conscious goal for some school systems. School systems in at least 13 states and the District of Columbia are seeking to "align" their curriculum so that students do not spend hours studying materials upon which they will never be tested regardless of the value or benefits which could be derived from that effort.[88] As a result, curriculum alignment "subordinates the process of curriculum development to external testing priorities. . . Thus, the curriculum falls in line with the test, and, for all intents and purposes, the test becomes the curriculum."[89] The psychometric approach, rooted in outmoded psychology, thus comes to control education.

The educational price paid for allowing tests to dictate the curriculum can be a high one. Julia R. Palmer, Executive Director of the American Reading Council, recently wrote, "[T]he major barrier to teaching reading in a common-sense and pleasurable way is the nationally normed standardized second grade reading test." Ms. Palmer explained that the test questions force teachers and students to focus on "reading readiness" exercises and workbooks in their early grades and not on reading. As a result, many students become disenchanted with reading because they rarely get a chance to participate in it or to read anything of real interest to them.[90]

Mathematics instruction has also been harmed by the emphasis on testing. Constance Kamii reports that the tests are unable to distinguish between students who understand underlying math concepts and those who are only able to perform procedures by rote and are thus unable to apply them to new situations. Focusing instruction on teaching to the test, therefore, precludes teaching so that children grasp the deeper logic.[91] The National Council of Teachers of Mathematics has concluded that unless assessment is changed, the teaching of math cannot improve.[92]

Unless assessment is changed, the teaching of mathematics cannot improve.

In general, test control over the curriculum narrows education, "dummies-down" learning materials and instruction, and makes school less interesting to both students and teachers.[93] This narrowing of curriculum is a virtually unavoidable by-product of emphasizing instruments of limited construct validity that utilize a multiple-choice format. As teaching becomes test coaching, real learning and real thinking are crowded out in too many schools.

Just as curricula have been narrowed, so too have textbooks. Diane Ravitch argues that "textbooks full of good literature began to disappear from American classrooms in the 1920's, when standardized tests were introduced. Appreciation of good literature gave way to emphasis on the 'mechanics' of reading."[94] Similarly, a recent report by the Council for Basic Education concluded that the emphasis on standardized tests and curriculum alignment are among the main causes of the increasingly poor quality of textbooks. The report noted that "instead of designing a book from the standpoint of its subject or its capacity to capture the children's imagination, editors are increasingly organizing elementary reading series around the content and time of standardized tests. . . As a result, much of what is in the textbooks is incomprehensible."[95] As Goodman, *et al.* point out, many textbooks have end-of-chapter or end-of-section exams that are badly made standardized, multiple-choice tests, and these exams define the content of the text and control how the material is taught.[96] These tests may be even more prevalent than regular standardized tests.[97]

By narrowing the curriculum, standardized tests are undermining many of the most important goals of the current school improvement movement.

Finally, by narrowing the curriculum, standardized tests are undermining many of the most important goals of the current school improvement movement. Recent education reform efforts have sought to promote "higher-order thinking skills," imagination and creativity in American students. Yet standardized tests focus on basic skills, not critical thinking, reasoning or problem-solving. They emphasize the quick recognition of isolated facts, not the more profound integration of information and generation of ideas.[98]

Several studies have demonstrated that "teaching behaviors that are effective in raising scores on tests of lower-level cognitive skills are nearly the opposite of those behaviors that are effective in developing complex cognitive learning, problem-solving ability, and creativity."[99] As Linda Darling-Hammond of the Rand Corporation concluded, "It's testing for the TV generation — superficial and passive. We don't ask if students can synthesize information, solve problems or think independently. We measure what they can recognize."[100]

Teachers, in general, are not pleased with the massive increases in testing.

Teachers, in general, are not pleased with the massive increases in testing. Research has shown that teachers believe that too much time is spent on testing and that tests don't measure student achievement very comprehensively, are not accurate for minority students, are not necessary for placement decisions, and are not instructionally useful. Further, most teachers believe that teaching to the test is educationally wrong, but has become necessary due to the pressure to raise test scores.[101]

Impact on Student Progress and Achievement

Standardized tests perpetuate and even exacerbate existing inequities in educational services, particularly for minority and low-income students.

By controlling or compelling student placement in various educational programs, standardized tests perpetuate and even exacerbate existing inequities in educational services, particularly for minority and low-income students. Thus, standardized test results lead to larger numbers of racial and ethnic minorities being placed in special education and remedial programs. For example, national data shows that African-American children are three times as likely as white children to be placed in special education programs.[102] A number of sources have reported to FairTest that school districts are placing ever more children in special education programs so that their test scores are not counted in the school or district averages.[103]

Standardized tests also perpetuate the domination of white upper-middle class students in "advanced" classes. In New York City for example, IQ tests are used in some districts to place children in "gifted and talented" programs, creating white, upper-middle class enclaves in districts whose enrollment is dominated by racial and ethnic minorities.[104] Similarly, test results assign boys to advanced math and science programs and keep girls out.

At the same time, standardized tests, particularly when used as promotional gates, can act as a powerful exclusionary device — again aimed disproportionately at minority and low-income students. In the end, they both narrow the educational opportunities available to many segments of our student population and maintain the isolation of different racial and social groups and classes.[105] For

example, academic research has demonstrated that, for a student who has repeated a grade, the probability of dropping out prior to graduation increases by 20 to 40%.[106] Thus, students who are not promoted because they have failed an often unreliable, invalid and biased standardized test are more likely to become high school dropouts.

Students who are not promoted because they have failed an often unreliable, invalid and biased standardized test are more likely to become high school dropouts.

The impact of standardized tests is particularly devastating when used to determine readiness for first grade. These tests are among the least valid and reliable and are among the most difficult to administer under relatively uniform conditions. Moreover, Shepard and Smith, after examining 14 controlled studies on the effects of kindergarten retention, concluded that retention provided no increase in subsequent academic achievement while imposing a significant social stigma on the retained students.[107]

Nor does the use of standardized tests affect only low-achieving students. High-achieving students are likely to be frustrated by a narrowed curriculum, which has been "dumbed down" in response to standardized tests, particularly minimum competency tests. These students, too, are likely to drop out.[108]

One of the most insidious effects of the overuse of standardized tests is on teachers' perceptions of their students. The existence of a "Pygmalion effect" as it relates to test results has long been a source of controversy. However, a 1984 study by Stephen Raudenbush has carefully documented its existence for students entering a new school (in this case, 7th grade students entering junior high school).[109] Where teachers have little information on students, conclusions about student knowledge, skills, and abilities based on often inaccurate and unreliable test results can become self-fulfilling prophesies.

Standardized tests are also used to determine pupil placements within regular programs, a practice often termed "tracking." Low-scoring students are placed in slower tracks while high-scoring students are placed in faster tracks. In the slower tracks, students receive a watered-down curriculum at a slower pace.[110] In lower tracks, students also are far more apt to receive instruction that focuses on rote and drill. Allington has documented the differences in reading instruction in various groups in which children in lower groups do not read much, and similar problems have been noted in other subject areas.[111]

Once tracked into a slow group, a student usually falls further and further behind those in higher groups, in terms both of content and skills learned and of test scores.

Once tracked into a slow group, a student usually falls further and further behind those in higher groups, in terms both of content

and skills learned and of test scores. This problem is even more acute for those labled "learning disabled." One consequence is that in the effort to increase test scores in the short run, students in lower tracks receive ever more instruction that resembles test-taking. That is, their schooling becomes reduced to test coaching.

The negative effect testing has on the curriculum damages low-income and minority students most of all.

Students in the slower tracks are disproportionately from low-income or minority-group backgrounds. This has led to segregation in many schools.[112] Minority and low-income students are administered biased and invalid tests, tracked into special education or slow programs, taught what often amounts to no more than test-preparation, and fall further behind their peers. The negative effect testing has on the curriculum damages low-income and minority students most of all. Their subsequent low test scores "confirm" the predictions made by earlier tests, and the growing test-score differentials are used to justify policies of tracking and retention.

Increasing evidence has shown that tracking, like retention, is rarely helpful. Students in the slower groups are clearly harmed by the process, but students in faster groups do not necessarily gain. Multi-ability classrooms would be preferable for low-achievers and not harm high achievers.[113]

Impact on Local Control

Because standardized tests increasingly determine what is taught in the classroom, parents and other citizens are losing their traditional control over the public schools. This shift of power from local communities to state and national government reduces the level of input and influence available to both parents and teachers in the management of the schools. This, in turn, reduces "the responsiveness of schools to their clientele and so reduces the quality of education" available in those schools.[114]

Local control over the schools is being lost to private organizations, namely the test developers.

Local control over the schools is also being lost to private organizations, namely the test developers. Because of the influence of testing on curricula and instruction, test-makers are effectively dictating the content and form of education. Test publishers and textbook publishers are often divisions of the same company, and both are increasingly owned by international corporations. One result is "*greater* centralization, tighter control ... [and] less accountability ... as the channels of intellectual thought are controlled by fewer and fewer publishers."[115]

Despite the significant and growing role their products play in educational decisions, test manufacturers face little government

regulation or supervision. Unlike other businesses, such as communications, food & drugs, transportation, and securities, there are virtually no regulatory structures at either the federal or state level governing the billion dollar a year testing industry.

States and school districts have neither the expertise nor the resources to independently develop and validate the standardized tests that they need. Instead they turn to private testing companies, who design and market a tremendous variety of products. Even here, states and school systems have neither the skills nor the funds to adequately investigate claims by test developers regarding test validation or to review the test validation process.[116]

There are virtually no regulatory structures at either the Federal or State level governing the billion dollar a year testing industry.

Even if the expertise and resources did exist, the secrecy which is rampant in the testing industry would likely prevent any effective outside evaluation. As the late Dr. Oscar K. Buros (editor of the *Mental Measurement Yearbook*) lamented, "It is practically impossible for a competent test technician or test consumer to make a thorough appraisal of the construction, validation, and use of standardized tests. . . because of the limited amount of trustworthy information supplied by the test publishers."[117]

"It is practically impossible for a competent test technician or test consumer to make a thorough appraisal ...because of the limited amount of trustworthy information supplied by the test publishers."
—*Oscar K. Buros*

Testing: An Invalid Enterprise

In sum, current standardized, multiple-choice tests are severely flawed instruments. Their overuse and misuse cause substantial individual and social harm. Many factors contribute to these problems:

- Test-makers make assumptions about human ability that cannot be proven but that lead directly to harmful educational assumptions and practices.

- No test is sufficiently reliable to be used as the sole or primary criteria for decision-making, but such decisions are made constantly.

- The content validity of tests is inadequate because the tests cannot measure the complex material contained in most learning or performance domains.

- Predictive criterion validity is too low to use tests as the sole or primary criteria for decision-making. The limited degree of validity that does exist often results from self-fulfilling prophecies.

- The construct validity of tests is also inadequate: tests often do not measure the traits they claim to measure or do so only poorly.

•Standardized exams often fail to accurately measure persons from atypical backgrounds, and test results are used to segregate and devalue persons from minority groups.

•The effects of testing not only cause irreparable harm to many individuals, they also are destructive to the educational process as a whole. Low-income and minority-group students are disproportionately subjected to the poorest, narrowest, most rigidly test-driven curriculum and instruction.

Most standardized, multiple-choice-type testing is invalid.

If, as some of testing's foremost theoreticians suggest, the validity of testing is inseparable from its social consequences, then most standardized, multiple-choice-type testing (including I.Q., readiness and most developmental screening testing) is invalid. Continued reliance on standardized testing will prevent necessary school reform. The ongoing domination of testing means that millions of students, predominantly those most in need of improved education, will be dumped into dead end tracks and pushed out of school. To prevent damage and to allow needed reforms, standardized testing must become an occasional adjunct, used for attaining basic but limited information about educational policies, and not a controlling factor over students or curricula.

IV. AN AGENDA FOR TESTING REFORM

FairTest concurs with the National Academy of Education that "information on student progress, wisely interpreted, is of obvious value to the public, to educators and to policy-makers at all levels of government."[118] If properly constructed, validated, administered and used, standardized tests could serve as one limited tool in this effort.

Unfortunately, it has become all too obvious that standardized tests are *not* properly constructed, validated and administered. Moreover, their widespread use is creating serious problems for students, teachers and the schools themselves. Reliance on standardized tests is now a roadblock to significant reform in curriculum and pedagogy. The question arises then: What should be done to reform tests and test use in the public schools?

In response to the misuse of standardized tests in U.S. society, FairTest has developed an agenda to answer this question. Our Testing Reform Agenda is guided by two essential principles: standardized, multiple-choice testing should be supplanted in most instances by authentic forms of evaluation, and standardized tests that remain must be markedly improved in their construction and use. The FairTest Reform Agenda has four parts:

New, authentic assessments must be used to measure and evaluate student achievement.

•It is time to use new methods of measuring student achievement as part of a major reform in schooling as a whole. Standardized multiple-choice tests can only measure a very limited range of student knowledge, abilities and skills. Current standardized tests should become no more than occasional complements to performance-based evaluations. Their use should largely be limited to matrix survey-samples, which will provide any programmatically useful information while limiting their effect on curriculum and preventing their being used to make decisions about individuals.

Emerging methods, commonly referred to as "authentic" and "performance-based" evaluation (See Appendix B, "What Is Authentic Evaluation?"), provide new opportunities to expand society's capability to more fully and accurately evaluate a greater range of knowledge, abilities and skills. Authentic evaluations can and should be used to diagnose the strengths and weaknesses of students in order to help them learn, rather than to sort, stratify or segregate them. Further, good assessment can encourage the teaching of challenging and comprehensive curricula in ways that spur

students to serious thinking and greater accomplishments. Today, "teaching to the test" is a synonym for narrowing education; with authentic assessments, "teaching to the test" can become a method of enriching and expanding education.

Testing should be undertaken when it can be directly helpful to student learning; any other reasons must be carefully justified and not allowed to negatively affect instruction. Tests and other forms of educational evaluation, whether performance-based or multiple-choice, should measure meaningful and important knowledge and capabilities possessed by students. Questions must be relevant to the knowledge, abilities or skills being tested. Test items and instructions should be written clearly and accurately. The tests themselves should take into account the diversity of language, experience and perspective embodied in the test-taking population. If necessary, different assessments should be used for different population groups to ensure the elimination of bias. At the same time, questions and scoring procedures should acknowledge the complexity and diversity of intelligence and individual development.

Tests must be properly constructed, validated and administered.

•Test validation should ensure that the content of the test matches the content of what is taught, but test developers cannot stop at content validation. They must document assumptions about the relationship between test results and future performance, and do so in ways that are more than documentation of a self-fulfilling prophecy. At the same time, they must demonstrate that test results are accurately related to the underlying knowledge, skills and abilities the test claims to measure, and that the theoretical conceptions of knowledge and learning used to develop test constructs are accurate. Test companies should not provide sub-scores that lack adequate reliability and validity, nor make recommendations for instructional practice when the tests have not been validated as diagnostic instruments.

Those who develop and use tests must ensure that the testing environment is both consistent for and supportive of all test-takers. Where the environment cannot be made standard and supportive, the only alternative is to refrain from testing. Moreover, the standard environment must not be constructed in a manner that creates disadvantages for particular students through artificial distractions or pressures.

Tests should be open.

•Public schools, test-takers and independent researchers should have access to the descriptive and statistical data needed to verify test publishers' claims regarding test construction and validation. This should include the release of questions used on previous tests,

as well as data on test results, identified by race/ethnicity, gender, socio-economic status, geographical residence, and other distinctions.

Test publishers have long argued against the release of old test questions. They have claimed that any large-scale release would require the development of a massive number of new items, thus increasing test development cost. This may not be the case. One study found that the release of old test questions did not affect test scores.[119] Nor has this been a problem in college admissions testing, where tests have been disclosed since 1980. Thus the release of old test questions may not require such a large scale development of new questions. On the other hand, some test publishers also publish test coaching books that contain material nearly identical to the tests. Some critics have called the use of these materials "cheating."[120] Test publishers should not sell coaching materials that undermine the validity of their tests.

Test users or independent public agencies should also fully investigate the claims of test publishers regarding the construction and validity of their tests. Test company materials as well as data to facilitate independent analysis must be made available by test companies. At the same time, test administrators and users should disclose and monitor their own processes for test administration.

Tests should be viewed in the proper perspective.

•Both test developers and test users should work to ensure that test results are properly interpreted and employed by educators, policymakers, test-takers and the general public. At a minimum, test scores should never be used as the sole or primary factor in "high stakes" educational decisions.

At the same time, test developers and test users must recognize that standardized tests are only limited measures of educational progress. Used alone, they present distorted pictures of what they seek to measure and their use often undermines the quality of education offered in our public schools, particularly the education offered to low-income and minority-group students. Both test developers and test users have the affirmative obligation to promote a proper, reasonable and limited use of standardized tests as one of a series of assessment mechanisms.

Although FairTest believes that those institutions that develop and use standardized tests have a primary obligation to reform tests and test use, government also has a major role to play. By establishing guidelines for the testing industry, requiring information on standardized tests to be made public, and analyzing test results to

guard against bias, the government can go a long way toward improving the quality of tests and test use. More importantly, public agencies can set the standard for intelligent and proper use of tests.

Unfortunately, too many policymakers and educators have ignored the complexities of testing issues and the obvious limitations they should place upon standardized test use. Instead, they have been seduced by the promise of simplicity and objectivity. The price which has been paid by our schools and our children for this infatuation with tests is high. Unless Americans act now to limit and reform the use of standardized tests in the schools, that price will continue to increase.

The price which has been paid by our schools and our children for this infatuation with tests is high.

NOTES

[1] Haney, W. and G. Madaus. "Effects of Standardized Testing and the Future of the National Assessment of Educational Progress," Working Paper for the NAEP Study Group (1986), p. 5.

[2] National Academy of Education. *The Nation's Report Card: Improving the Assessment of Student Achievement, Review of the Alexander/James Study Group Report* (Cambridge, MA: National Academy of Education, 1987) p. 47.

[3] Wigdor, A.K. and W.R. Garner, eds. *Ability Testing: Uses, Consequences & Controversies* (National Academy Press, 1987) p. 47.

[4] Education Research Service, Inc. *Kindergarten Programs & Practices in Public Schools* (Education Research Service, Inc., 1986).

[5] Peterson, B. "Half Million Standardized Tests Given to MPS Students," *Rethinking Schools* (May-June, 1989), and personal communication from Peterson to author Neill. Limited-English Proficiency students were not included in the total. The Task Force also counted 860,000 end-of-unit basal reader tests; if these are included as standardized tests, then Milwaukee administers nearly 1,350,000 tests per year, or about 14 per student per year. As K. Goodman, *et al.*, report, tests in basal readers are, in general, less well developed but essentially the same thing as standardized multiple-choice achievement tests (*Report Card on Basal Readers*, Katonah, NY: Richard C. Owen Publishers, 1988).

[6] Pipho, C. "Tracking the Reforms: Part 5 - Testing." *Education Week* (May 22, 1985) p. 19.

[7] Southern states may be taking the lead in reducing the amount of standardized testing. Over the past several years, North Carolina banned the use of standardized achievement tests in grades one and two and is implementing developmentally appropriate assessments in those grades; Mississippi stopped requiring the testing of young children; Georgia dropped, after one year, a required test for admission to first grade; Kentucky has lifted its requirement for annual testing in every grade and may be turning to performance-based assessments (see Appendix B for discussion of performance-based assessment).

[8] Since this was written in 1988, Ohio has also adopted state-wide testing. Vermont is considering performance-based assessments through portfolios, but may also use other methods of testing on a statewide basis.

[9] Congressional Budget Office. *Trends in Educational Achievement* (Washington, D.C.: Government Printing Office, April 1986) p. 20.

[10] Gannon, J. "College Grades and LSAT Scores: An Opportunity to Examine the 'Real Differences' in Minority-Nonminority Performance," in D. White, ed., *Towards a Diversified Legal Profession: An Inquiry into the Law School Admission Test, Grade Inflation, and Current Admissions Policies* (San Francisco: Julian Richardson Associates, 1981) p. 273.

[11] Ames, L.B., *et al. The Gesell Institute's Child from One to Six* (New York: Harper & Row, 1979) p. 168; Gesell instruments are constructed so that a "normal" boy just five years old shows up as developmentally four years and six months, effectively defining a large percentage of boys as developmentally unready. Rosser, P. *Sex Bias in College Admissions Tests: Why Women Lose Out* (Cambridge, MA: FairTest, 3rd Ed., 1989).

[12] Hoover M.R., R.L. Politzer and O. Taylor. "Bias in Reading Tests for Black Language Speakers: A Sociolinguistic Perspective," *Negro Educational Review* (April-July, 1987) pp. 81-98.

[13] Loewen, J. "Possible Causes of Lower Black Scores on Aptitude Tests." (Burlington: University of Vermont, unpublished research report, 1980).

[14] Taylor, O.L. and D.L. Lee. "Standardized Tests and African-American Children: Communication and Language Issues," *Negro Educational Review* (April-July, 1987) pp. 67-80.

[15] Meier, T. "The Case Against Standardized Achievement Tests," *Rethinking Schools* (Vol. 3, No. 2, 1989) p. 12. S.B. Heath. *Ways With Words: Language, Life, and Work in Communities and Classrooms* (New York: Cambridge University, 1983), cited in Meier (1989).

[16] Butler, J. "Looking Backward: Intelligence and Testing in the Year 2000," *National Elementary Principal* (March-April, 1975) p. 74.

[17] For research on Hispanics, see A.P. Schmitt. "Unexpected Differential Item Performance of Hispanic Examinees on the SAT-Verbal, Forms 3FSAO8 and 3GSAO8" (Princeton, NJ: Educational Testing Service, Unpublished statistical report, 1986). Dr. Schmitt, an ETS researcher, concluded that Mexican-American students scored significantly higher than expected on a reading comprehension passage concerned with lifestyle changes in Mexican-American families. For research on Blacks, see Hoover, Politzer & Taylor (1987) p. 83, who note that Dr. Darlene Williams found that "the use of pictures showing Blacks and related to Black culture raised IQ scores for all Black children." For research on females, see J.W. Loewen, P. Rosser & J. Katzman. "Gender Bias in SAT Items." (Paper presented at the American Education Research Association Annual Conference New Orleans, La., April 5, 1988).

[18] The mathematics section of the WISC-R test, for example, includes 8 questions about 13 boys or men who save money on purchases, trade fairly, cleverly divide their efforts and money, and work at jobs, compared to only one question featuring a girl—who loses her hair ribbon. Wechsler, D. *Wechsler Intelligence Scale for Children-Revised* (Psychological Corporation, 1974).

[19] Scheuneman, J.D. "A Posteriori Analyses of Biased Items," in R.A. Berk, ed.,*Handbook of Methods for Detecting Test Bias* (Baltimore: Johns Hopkins, 1982). L.A. Shepard. "Identifying Bias in Test Items," in B.F. Green, ed., *New Directions for Testing and Measurement: Issues in Testing - Coaching, Disclosure and Ethnic Bias*, No. 11 (San Francisco: Josey-Bass, 1981).

[20] Shepard (1981). Berk, ed., (1982), Ch. 9, "Methods Used by Test Publishers to 'Debias' Standardized Tests." Angoff, W.H. "Philosophical Issues of Current Interest to Measurement Theorists," Educational Testing Service Research Report RR-87-33 (Princeton, N.J.: Educational Testing Service, August 1987).

[21] Medley, D.M. and T.J. Quirk. "The Application of a Factorial Design to the Study of Cultural Bias in General Culture Items on the National Teacher Examination," *Journal of Educational Measurement* (Vol. II, No. 4, Winter, 1974). Hackett, R.K., *et al.*, "Test Construction Manipulating Score Differences Between Black and White Examinees: Properties of the Resulting Tests" (Princeton, N.J.: Educational Testing Service, 1987). *Educational Measurement*, "Golden Rule or Golden Ruse" (Summer 1987), special section, containing six articles on the Golden Rule settlement and the Golden Rule bias reduction procedure.

[22] Hoffman, B. *The Tyranny of Testing* (New York: Crowell-Collier, 1962) pp. 60-61.

[23] Gould, S.J. *The Mismeasure of Man* (New York: Norton, 1981) p. 24.

[24] Levidow, L. "'Ability' Labeling as Racism," in Levidow, L. and D. Gill, *Anti-Racist Science Teaching* (London: Free Association Books, 1987) p. 239. He also observes that "the 'intelligence' of IQ testing is constructed by the testing process itself" (p. 235).

[25] Singh, B. "Graded Assessments," in Gill and Levidow, eds., (1987).

[26] National Association for the Education of Young Children. "NAEYC Position Statement on Developmentally Appropriate Practice in the Primary Grades, Serving 5- Through 8-Year-Olds," *Young Children* (January 1988).

[27] Taylor and Lee (1987).

[28] Meier, D. "Why Reading Tests Don't Test Reading," *Dissent* (Winter, 1982-83). Meier notes that questions correctly answered by "low-performing" students, but not by "high-performing" students are removed from test use as inaccurate measures of ability or achievement. This ignores the possibility that these questions reflect differing language or cognitive styles possessed by "low-performing" students. See also A. Martin "Screening, Early Intervention, and Remediation: Obscuring Children's Potential," *Harvard Educational Review* (Vol. 58, No. 4, Nov., 1988) pp. 488-501; in which she points out that the testing and screening processes produce an overemphasis on children's weaknesses, treat variation as a problem, and disempower teachers.

[29] Haney, W. "Test Reasoning and Reasoning about Testing," *Review of Educational Research* (Winter, 1984) p. 635. Quoting O.K. Buros.

[30] Witt. J.C. and F.M. Gresham. "Review of WISC-R," *Ninth Mental Measurement Yearbook* (1985) p. 1716. The WISC-R itself has not changed since 1974. For readiness testing, see S. Meisels, "Uses and Abuses of Developmental Screening and School Readiness Testing," *Young Children* (January 1987).

[31] Resnick, L.B. and D.P. Resnick, "Assessing the Thinking Curriculum: New Tools for Educational Reform," in B.R. Gifford and M.C. O'Connor, eds., *Future Assessments: Changing Views of Aptitude, Achievement, and Instruction* (Boston: Kluwer Academic Publishers, 1989).

[32] Ryan, C. *The Testing Maze* (Chicago, Ill.: National PTA, 1979) p. 8. S.A. Cohen. "Instructional Alignment," *Educational Researcher* (November, 1987).

[33] Gardner, H. *Frames of Mind: The Theory of Multiple Intelligences* (New York: Basic Books, 1985).

[34] Hoffman (1962) pp. 54-56. Meier (1982-3).

[35] Kaufman, N.L. "Review of Gesell Preschool Test," *Ninth Mental Measurement Yearbook* (1985) p. 607.

[36] Tittle, C.K. "Review of WISC-R," *Eighth Mental Measurement Yearbook* (1978) p. 353.

[37] *Ninth Mental Measurement Yearbook* (1985). See reviews of the California Achievement Test, Comprehensive Tests of Basic Skills, Iowa Tests of Basic Skills, Metropolitan Achievement Test, Stanford Achievement Test, and SRA Achievement Series.

[38] Cannell, J.J. *Nationally Normed Elementary Achievement Testing in America's Public Schools* (Daniels, WV: Friends for Education, 1987) p. 6. This report examined the same six tests listed in the previous footnote.

[39] Koretz, D. "Arriving in Lake Wobegon," *American Educator* (Summer, 1988) p. 8-52.

[40] Rothman, R. "Physician's Test Study was 'Clearly Right,'" *Education Week* (April 5, 1989).

[41] Anastasi, A. *Psychological Testing* (New York: Macmillan Publishing Company, 6th Edition, 1988). See discussions in Chapters 3 and 5.

[42] Anastasi (1988). See also reviews in *Ninth Mental Measurement Yearbook* (1985).

[43] *Ninth Mental Measurement Yearbook* (1985). See reviews of tests listed in footnote 37.

[44] Anastasi (1988). See discussion in Chapter 5.

[45] Cited in Shepard, L.S. & M.L. Smith. "Flunking Kindergarten: Escalating Curriculum Leaves Many Behind," *American Educator* (Summer, 1988) p. 36.

[46] Kaufman (1985) p. 607.

[47] College Entrance Examination Board. *1989-90 ATP Guide for High Schools and Colleges* (New York: CEEB, 1989).

[48] *Ninth Mental Measurement Yearbook* (1985). See reviews of tests listed in footnote 37.

[49] Congressional Budget Office (1986) p. 10. A second report by CBO. *Educational Achievement: Explanations and Implications of Recent Trends* (Washington, D.C.: Government Printing Office, August, 1987) p. 10-11, concluded that "one of the most serious mistakes made by some analysts attempting to explain recent achievement trends . . . has been to assume that patterns evident in the scores of one test will appear in others as well."

[50] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing* (Washington, D.C.: APA, 1985), Standard 8.12 (p.54): "In elementary or secondary education, a decision or characterization that will have a major impact on a test taker should not automatically be made on the basis of a single test score."

[51] Anastasi (1986) pp. 34, 38.

[52] Wodtke, K., F. Harper, M. Schommer and P. Brunelli. "Social Context Effects in Early School Testing: An Observational Study of the Testing Process" (Paper for the American Educational Research Association, 1985) p. 28.

[53] Fuchs, D. and L.S. Fuchs. "Test Procedure Bias: A Meta-Analysis of Examiner Familiarity Effects," *Review of Educational Research* (Summer, 1986) pp. 243-262. See also "Test Conditions Can Harm Minority-Group Children," *The Chronicle of Higher Education* (November 18, 1987) p. A15.

[54] Maeroff, G.I. "Reading Test Time Limits Are Criticized," *New York Times* (January 29, 1985) p. C-1. Dorans, N.J., *et al.*, "Differential Speededness: Some Items Have DIF Because of Where They Are, Not What They Are" (Paper for the National Council on Measurement in Education, April 1988). C.A. Dwyer, "Test Content in Mathematics and Science: The Consideration of Sex" (Paper for the American Educational Research Association (AERA), April 1976). A.P. Schmitt, "Language and Cultural Characteristics That Explain Differential Item Functioning for Hispanic Examinees on the Scholastic Aptitude Test," *Journal of Educational Measurement* (Spring 1988). See also, Anastasi (1988) p. 35 for additional examples of examiners' effects on test scores. And see C.L.W. Wendler and S.T. Carlton, "An Examination of SAT Verbal Items for Differential Performance by Women and Men: An Exploratory Study" (Paper for the AERA, April 1987) indicating that unwillingness to guess may also lower women's scores.

[55] Airasian, P.W. and G.F. Madaus. "Linking Testing and Instruction: Policy Issues," *Journal of Educational Measurement* (Summer, 1983) p. 104.

[56] Madaus, G. and D. Pullin. "Questions to Ask When Evaluating a High-Stakes Testing Program," *NCAS Backgrounder* (National Coalition of Advocates for Students, June 1987). See also Anastasi (1988), Chapter 6.

[57] For discussions of this problem as regards the NTE (National Teacher Exam), see: Horner, B. and J. Sammons. *The Test That Fails: An Analysis of the National Teachers Examination in New York* (New York: NYPIRG, 1987) pp. 4-6; L. Darling-Hammond. "Teaching Knowledge: How Do We Test It?" *American Educator* (Fall, 1986) p. 88. FairTest recently received a questionnaire circulated by ETS asking respondents to indicate whether certain content areas ought to be on the NTE. Not only did the questionnaire demonstrate continued use of the affirmative

hypothesis, but also it demonstrated continuing reliance on using separable bits of "knowledge" that can be reduced to multiple-choice questions.

[58] Madaus and Pullin (1987).

[59] Airasian, P.W. "Review of Iowa Tests of Basic Skills," *Ninth Mental Measurement Yearbook* (1985) p. 719.

[60] *Ninth Mental Measurement Yearbook* (1985). See reviews of tests listed in footnote 37.

[61] Mitchell, B.B. "Predictive Validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis for White and for Negro Pupils," *Educational and Psychological Measurement* (1967) pp. 1047-1054. See also, P.H. Johnston, "Chapter 6: Assessment in Reading," in P.D. Pearson, ed., *Handbook of Reading Research* (1984) p. 162. The tendency is for test-maker's evidence on criterion-related validity to take the form of "Test A is valid because test B is valid because test C is valid, etc."

[62] Congressional Budget Office (1986) p. 20.

[63] Johnston, P. "Teachers as Evaluation Experts," *The Reading Teacher* (April, 1987) pp. 744-748. P. Johnston (1984).

[64] Shepard and Smith (1988). See also, "Mass Academic Testing of Young Children Should Stop, Groups Argue," *Education Week* (March 25, 1988) p. 5.

[65] Meisels, S.J. *Developmental Screening in Early Childhood: A Guide,* (Washington, D.C.: National Association for the Education of Young Children, Third Edition, 1989).

[66] Meier, D. (Winter, 1982-83). See also, A. Bussis. "'Burn It at the Casket': Research, Reading Instruction, and Children's Learning of the First R," *Phi Delta Kappan* (December, 1982); A. Bussis and E.A. Chittendon. "What the Reading Tests Neglect," *Language Arts* (March, 1987); E.A. Chittendon. "Styles, Reading Strategies and Test Performance: A Follow-Up Study of Beginning Readers," in R.O. Freedle and R.P. Duran, eds., *Cognitive and Linguistic Analyses of Test Performance* (Ablex, 1987); C. Edelsky, and S. Harman, "One More Critique of Reading Tests - With Two Differences," *English Education* (Oct., 1988).

[67] Harste, J. in *Reading Today* (Dec., 1989 - Jan., 1990).

[68] Archbald, D.A., and F.M. Newmann. *Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School* (National Association of Secondary School Principals, 1988); C. Kamii. *Young Children Continue to Reinvent Arithmetic, 2nd Grade* (Teachers College Press, 1989); E. Pardo, and J.C. Russell. "Standardized Tests at the Early Childhood Level: Do They Tell Us the Truth about Children?" (Paper for the American Educational Research Association, March 1989); Resnick and Resnick (1989). Shepard L., "Why We Need Better Assessments," *Educational Leadership* (April, 1989); Smith, F. *Insult to Intelligence* (Arbor House, 1986); Smith, M.L. *et al., The Role of Testing in Elementary Schools* (Tempe, AZ: Arizona State University, Department of Education, Unpublished Monograph, 1990).

[69] Airasian and Madaus (1983).

[70] Messick, S. "Meaning and Values in Test Validation," *Educational Researcher* (March, 1989) pp. 5-11; Messick, S. "The Once and Future Issues of Validity," in H. Wainer and H. Braun. *Test Validity* (Hillsdale, N.J.: Lawrence Erlbaum, 1988) pp. 33-45. Cronbach, L. "Five Perspectives on the Validity Argument," in Wainer and Braun (1988) pp. 3-18.

[71] Messick (1989, 1988). See also: Cronbach (1988); C.K. Tittle. "Validity: Whose Construction Is It in the Teaching and Learning Context?," *Educational Measurement* (Spring, 1989) pp. 5 - 13; R.E. Schutz. "Faces of Validity of Educational Tests," *Educational Evaluation and Policy Analysis* (Summer, 1985) pp. 139-142.

[72] Messick (1989).

[73] Neill, D.M. "Standardized Testing: Harmful to Civil Rights," (Washington, DC: United States Commission on Civil Rights, forthcoming).

[74] Frederiksen, N. "The Real Test Bias: Influences of Testing on Teaching and Learning," *American Psychologist* (March, 1984) pp. 193-202. R. Marzano and A. Costa. "Question: Do Standardized Tests Measure General Cognitive Skills? Answer: No," *Educational Leadership* (May, 1988) pp. 66-71.

[75] Frederiksen (1984); Kamii (1989); Meier, D. (1982-3); Oakes, J. (1985); and M. Lipton, "Examining Curriculum in the 'Best' Schools," *Education Week* (March 7, 1990) p. 36.

[76] Resnick and Resnick (1989).

[77] Johnston, P. "Constructive Evaluation and the Improvement of Teaching and Learning," *Teachers College Record* (Summer, 1989) pp. 509-528.

[78] Dilworth, M.E. *Teachers' Totter: A Report on Teacher Certification Issues* (Institute for the Study of Educational Policy, 1984) pp. 33-34.

[79] National Association for the Education of Young Children. "NAEYC Position Statement on Standardized Testing of Young Children 3 Through 8 Years of Age," *Young Children* (1988) pp. 42-47, and *Testing of Young Children: Concerns and Cautions* (Washington, D.C.: NAEYC, pamphlet, 1988). For a general discussion of the harmful effects of achievement testing on young children, see C. Kamii, ed., *Achievement Testing in the Early Grades: The Games Grown-Ups Play* (Washington, D.C.: NAEYC, 1990).

[80] Madaus, G.F. "The Influence of Testing on the Curriculum," *87th Yearbook of the National Society for the Study of Education, Part I: Critical Issues in the Curriculum* (1988) pp. 83-121. See also, H.C. Rudman. "Testing Beyond Minimums," *ASAP Notes* (Occasional Paper # 5, 1985) pp. 1-36.

[81] Madaus, G.F. "Test Scores as Administrative Mechanisms in Educational Policy," *Phi Delta Kappan* (May, 1985) p. 616.

[82] "Some 'Teach' to the Test," *The Daily Press* [Newport News, VA] (June 15, 1987) p. C1.

[83] Haney and Madaus (1986) p. 19.

[84] Wise, A.E. "Legislated Learning Revisited," *Phi Delta Kappan* (January, 1988) p. 330. See also, D.W. Dorr-Bremme and J.L. Herman. *Assessing Student Achievement: A Profile of Classroom Practices* (Los Angeles, CA: Center for the Study of Evaluation, UCLA, 1986) which concluded, based upon a nationally representative sample of 114 school districts, that "preparations for a test can begin days or even weeks before the test is given." They also observed that "for each hour that students spend taking tests, teachers seem to spend two to three more" hours in preparation.

[85] Oral communication to author Neill (Jan., 1990).

[86] Smith, M.L. *et al.* (1990).

[87] Bishop, C.D. Presentation to the Joint Legislative Committee on Goals for Educational Excellence (Arizona Department of Education, Nov. 9, 1989).

[88] Olson, L. "Districts Turn to Nonprofit Group for Help in 'Realigning' Curricula to Parallel Tests," *Education Week* (October 18, 1987) pp. 1 & 19.

[89] Hlebowitsch, P.S. Letter to the editor, *Education Week* (November 18, 1987) p. 21.

[90] Palmer, J.R. Letter to the editor, *New York Times* (December 14, 1987). See also, J.T. Guthrie *Indicators of Reading Education* (New Brunswick, NJ: Center for Policy Research in Education, 1988), which concludes that the strengthening of students' reading skills goes hand-in-hand with finding better ways to measure reading achievement. The primary shortcoming of reading tests is that they don't reflect the complexity of the reading process. See also Note 66.

[91] Kamii (1989).

[92] National Council of Teachers of Mathematics. *Curriculum and Evaluation Standards for School Mathematics* (forthcoming). See also, National Research Council of the National Academy of Sciences. *Everybody Counts - A Report to the Nation on the Future of Mathematics Education* (Washington, D.C.: National Academy Press, 1989).

[93] McNeil, L.M. "Contradictions of Reform," *Phi Delta Kappan* (March, 1988) pp. 478-485.

[94] Fiske, E.B. "America's Test Mania," *New York Times* (April 10, 1988), Section 12, p. 20.

[95] Rothman, R. "Textbook Rules Have Backfired, Report Contends," *Education Week* (April 20, 1988) p. 1. See Tyson-Bernstein in Bibliography. See also K.S. Goodman, *et al. Report Card on Basal Readers* (Katonah, New York: Robert C. Owen Publishers, 1988), prepared for the National Council of Teachers of English.

[96] Goodman, K.S., *et al.* (1988).

[97] See note 5.

[98] Bastian, A. *et al., Choosing Equality: The Case for Democratic Schooling* (Philadelphia: Temple University Press, 1986) p. 73. See also Frederiksen (1984).

[99] McClellan, M.C. "Testing and Reform," *Phi Delta Kappan* (June, 1988) p. 769.

[100] Fiske (1988) p. 20.

[101] Ginsberg, R. and B. Berry. "Experiencing School Reform: The View from South Carolina," *Phi Delta Kappan* (March, 1990) pp. 549-552; McNeil (1988); M. L. Smith, *et al.*, (1990); Winthrop Rockefeller Foundation. *Fulfilling the Promises of Reform: Arkansas School Reform Study, 1985-1988* (Little Rock, AR: Author, 1988) pp. 22-30.

[102] Finn, J.D. "Patterns in Special Education Placement as Revealed by the OCR Surveys," in K. Heller, W. Holtzman and S. Messick, eds., *Placing Children in Special Education* (Washington, D.C.: National Academy Press, 1982). Because of test bias, a federal court in California has banned the use of IQ tests in the placement of Black children in classes for the "educable mentally retarded" (*FairTest Examiner*, Vol. 2, No. 1, 1988) p. 4.

[103] The city of Boston, for example, does not test some 30 percent of its students. City Wide Education Coalition "CWEC Fact Sheet on Boston Public Schools and Student Achievement" (Boston: CWEC, Nov., 1986).

[104] Cook, A. of Community Studies, Inc., New York City, oral communication (April 1988).

[105] National Coalition of Advocates for Students. *Barriers to Excellence: Our Children at Risk* (Boston: National Coalition of Advocates for Students, 1985). D. Pullin, "Educational Testing: Impact on Children at Risk" *NCAS Backgrounder* (Boston: National Coalition of Advocates for Students, 1985).

[106] Massachusetts Advocacy Center. "Memorandum to the Boston School Committee"(June 19, 1987) quoting from Office of Educational Assessment, New York City Board of Education. "Evaluation Update on the Effect of the Promotional Policy Program," (November 12, 1986). See also, L.A. Shepard and M.L. Smith. eds., *Flunking Grades: Research and Policies on Retention* (Falmer Press, 1989); M.L. Smith and L.A. Shepard. "What Doesn't Work: Explaining Policies of Retention in the Early Grades," *Phi Delta Kappan* (October, 1987) pp. 129-134.

[107] Shepard and Smith (1988) p. 34. See also, Shepard and Smith (1989) and Smith & Shepard (1987).

[108] "Student Competency Exams Present Major Barrier to Minority Students," *Education Daily* (August 27, 1987) p.3.

[109] Raudenbush, S. "Magnitude of Teacher Expectancy Effects on Pupil I.Q. As a Function of the Credibility of Expectancy Induction — A Synthesis of Findings From 18 Experiments," *Journal of Educational Psychology* (Vol. 76, No. 1, 1984) pp. 85-97.

[110] Oakes (1985).

[111] Allington, R.L. "Poor Readers Don't Get to Read Much in Reading Groups," Language Arts (Nov.-Dec., 1980) pp. 872-876; "The Reading Instruction Provided Readers of Different Abilities," *The Elementary School Journal* (Vol. 83, No. 5) pp. 548-559. "Shattered Hopes: Why Two Federal Reading Programs Have Failed to Correct Reading Failure," *Learning* (July-Aug., 1987) pp. 60-64. See also, A.N. Applebee, *et al., Crossroads in American Education: A Summary of Findings from NAEP Report Cards* (Princeton, NJ: Educational Testing Service, 1989).

[112] Chunn, E.W. "Sorting Black Students for Success and Failure: The Inequity of Ability Grouping and Tracking," pp. 93-106; J. L. Epstein, "After the Bus Arrives: Resegregation in Desegregated Schools," *Journal of Social Issues* (Vol 41, No. 3, 1985) pp. 23-43. Oakes (1985).

[113] Allington, (1980, 1983, 1987); Applebee, *et al.*, (1989); Oakes (1985).

[114] Wise (1988) pp. 328-333. See also, A. Porter, "Indicators: Objective Data or Political Tool?" *Phi Delta Kappan* (March, 1988) pp. 503-508.

[115] Rudman, H.C. "Corporate Mergers in the Publishing Industry: Helpful or Intrusive?" *Educational Researcher* (Jan. 1990) pp. 14-20.

[116] Madaus and Pullin, (1987) p. 3-4.

[117] Buros, O. "Fifty Years in Testing: Some Reminiscences, Criticisms, and Suggestions," *Educational Researcher* (July-August, 1977) p. 14.

[118] National Academy of Education (1987) p. 47.

[119] Haney (1984) p. 628.

[120] Mehrens, W.A. and J. Kaminski. "Methods for Improving Standardized Test Scores: Fruitful, Fruitless, or Fraudulent?" *Educational Measurement* (Spring, 1989) pp. 14-22.

APPENDIX A

DESCRIPTION OF FAIRTEST SURVEY AND RESULTS

Study Methodology

In mid-1987, FairTest staff conducted a series of telephone interviews with officials from all 50 state departments of education, from the District of Columbia school district and from 56 sample school districts in 38 states. Interviews of state officials focused on standardized tests administered by the public schools to fulfill testing mandates established by the state, while interviews of school district officials focused on tests administered to fulfill testing mandates established by the district. All interviews sought responses to three basic questions:

- How many tests were administered by the public schools to fulfill the state or local testing mandates?
- Which standardized tests were used to fulfill the state or local testing mandates?
- For what purposes did the state or school district mandate standardized tests?

The responses collected through these interviews related entirely to the use of standardized achievement, competency or basic skills tests. Public schools also use many other standardized exams, including IQ tests, behavioral tests, readiness tests for young children, and placement tests for special education, remedial education and bilingual education programs. However, the use of these tests varies considerably among schools, even within the same districts, and records of their use appear to be unreliable or nonexistent[1]. Thus the study results reflect only a portion of the standardized tests actually administered to students by the public schools.

Study Results

State-Level Testing Mandates. During the 1986-87 school year, schools in 42 states and the District of Columbia administered nearly 17 million standardized achievement, competency and basic skills tests to almost 36.3 million students to fulfill state testing mandates —a rate of almost one test for every two students[2]. FairTest counted each test battery as one test administration. (If individual tests within test batteries are counted, this number would easily double,

to 34 million tests.) This rate varied considerably among the states however. A detailed listing of the number of standardized tests administered by the public schools in each state to fulfill state testing mandates is presented in Table 1.

On average, schools in the South administered standardized tests to fulfill state mandates at much higher rates than schools in the remainder of the nation. Schools in the 11 Southern states (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee and Virginia) administered more than 6 million tests to over 9.3 million students — a rate of one test for every 1.5 students. Schools in the remaining 31 states administered tests at about half that rate — one test for every 2.5 students. In fact, schools in all but two Southern states (Florida and Virginia) administered tests at a rate higher than the average for the remainder of the nation. Moreover, the four states with the highest rates of test administration were all in the South. Kentucky, North Carolina, Alabama, and Georgia all administered about one test per student per year.

Outside the South, no clear patterns of test use to fulfill state mandates emerged. Although schools in New Mexico, a state with a minority population in the public schools above the national average, reported a high rate of standardized test use, so did schools in Utah, a state with a relatively low minority population. Conversely, schools in states like Texas and California, with relatively high minority populations in the public schools, reported rates almost equal to Wisconsin and Kansas, states with relatively low minority populations.

A clear pattern did emerge among the eight states (Alaska, Iowa, Minnesota, Montana, North Dakota, Ohio, Vermont and Wyoming) which did not have any state testing mandates. Seven of these eight states have minority student populations significantly below the national average. While minorities make up more than one-quarter of public school students nationwide, the minority student populations in these seven states range from a high of 15% in Ohio to a low of 1% in Vermont. Most in fact have minority populations that are less than 10% of the total student population. The one exception is Alaska, with a large Native American population, and an overall minority population just above the national average. However, Alaska will also implement a state testing mandate beginning in the 1987-88 school year.

District-Level Testing

During the 1986-87 school year, schools in the 56 sample school districts administered more than 7.8 million standardized tests to over 5.7 million students to meet local testing mandates — a rate of one and one-third tests for every student. (Again, FairTest counted only whole batteries, not tests within batteries.) However, the rate of test administration among districts varies even more than it did among different states. A detailed listing of the number of standardized tests administered by the public schools in each of the sample school districts to fulfill local testing mandates is presented in Table 2.

The schools in Newark, New Jersey are excluded from the discussion in the following paragraph due to their extremely high reliance on standardized tests. During the 1986-87 school year, schools in Newark administered over 600,000 standardized tests to about 67,000 students — a rate of more than 9 tests per student. This rate was more than three times that of the next highest school system.[3]

Excluding the Newark school system, the rate of test administration ranges from a high of almost 3 tests per student in Cleveland to a low of only one test per 12.5 students in Fairfax County, Virginia. Overall, six districts reported administering more than 2 tests per students per year, while seven others reported administering less than 1 test for every 2 students.

To analyze the variations among the different rates of tests administered by the different districts, districts were categorized by size. Three categories were created: large districts (with student populations exceeding 100,000); medium-sized districts (with populations between 25,000 and 100,000); and small districts (with populations below 25,000). The results from Newark were excluded from these calculations. On average, large districts administered standardized tests to fulfill local testing mandates at a rate (1.38 tests per student) 25% higher than that of medium-sized districts (1.11 tests per student). The rate of medium-sized districts, in turn, exceeded that of small districts (0.88 tests per student) by almost the same proportion.

The estimate that schools administered almost 38.9 million standardized achievement, competency and basic skills tests to 39.8 million students during the 1986-87 school year to fulfill local testing mandates is based upon these average variations and the distribution of students among school districts of different sizes.

(See Table 3 for the detailed computations.) Combining this figure with the tests administered to fulfill state testing mandates produces an estimate of 55.7 million standardized achievement-type tests administered by the public schools during the 1986-87 school year, or 1.4 tests for every public school student. (If we counted each test, not each battery, the total could approach three per student per year.)

Additional Surveys on Test Use

Although the FairTest survey focused on the use of standardized achievement, competency and basic skills tests in the public schools, public schools also use standardized tests for a variety of other purposes. These include:

- screening and readiness tests administered to kindergarten and pre-kindergarten students;

- tests administered to gifted, disadvantaged, handicapped and limited-English proficient students for placement into or graduation from gifted & talented, compensatory education, special education and bilingual education programs;

- tests administered to randomly selected samples of students as part of the U.S. Department of Education's National Assessment of Educational Progress.

- admissions tests administered to students seeking to enroll in particular secondary schools;

- college admissions tests administered to high school juniors and seniors;

- GED (General Education Development) tests administered to individuals who did not complete high school;

In addition, some school systems continue to administer IQ tests to their entire student populations, although most school districts administer IQ tests only to some individuals.

As noted previously, information on the number of other standardized tests administered was neither as specific nor as reliable as the data gathered for achievement, competency and basic skills tests. As a result, it is not possible to compute specific totals on the use of these tests. From a variety of sources, however, general figures can be obtained:

• Test publishers have reported that the total number of college and secondary school admissions, GED and NAEP tests administered to elementary and secondary school students was between 6 and 7 million.

• A survey conducted in the early 1980's indicated that students in compensatory education and special education programs were tested two to three times as often as their peers in mainstream education program.[4] Since the mainstream student is administered 1.4 standardized tests per year (according to our survey), compensatory and special education students are administered between 3 and 4 standardized tests per year. Given that over 10 million students participate in federally-funded compensatory and special education programs, an estimated 30 to 40 million additional standardized tests are administered to these students. (If individual tests rather than tests within batteries are counted, the numbers could double to 6-8 tests per students per year for a total of 60-80 million tests.)

• A survey conducted in 1985 concluded that almost half of the kindergarten and pre-kindergarten students in the public schools were administered screening tests.[5] Based upon the 1985 kindergarten enrollment, this means that between 1.5 and 1.75 million tests were administered to these children.

This total (which still excludes tests administered to gifted and limited-English proficient students and some proportion of I.Q. testing) yields an additional 37.5 to 48.75 million tests.

Summary

The results of the FairTest survey revealed that at least 55.7 million standardized achievement-type tests were administered to public school students in the 1986-87 school year. An additional 37.5 to 48.75 million standardized tests were administered by the public schools to their students for other purposes during that year. Based on these two estimates, between 93 million and 105 million standardized tests were administered to 39.8 million students in the public schools. If individual tests within batteries are counted, the total would increase dramatically. Looked at in this way, and including the uncounted IQ and other tests, the number of standardized tests administered yearly could approach 200 million, close to 5 per year per student.

NOTES ON SURVEY

[1] Harris J., and J. Sammons. *Failing our Children* (New York: New York Public Interest Research Group, 1989). This survey revealed that tests validated for one purpose are often used for other purposes for which the publisher claims no validity; e.g., achievement tests may be used to determine school readiness or even to screen for disabilities.

[2] See Chapter I, "Test Use in U.S. Schools," for some changes that have occured since 1987.

[3] In fact, recent evidence indicates that Newark may not be unusual—see Note 5 to the main text.

[4] Wigdor, A.K. and W.R. Garner eds. *Ability Testing: Uses, Consequences and Controversies* (Washington, DC: National Academy Press, 1982) pp. 252-253.

[5] Educational Research Services, Inc. *Kindergarten Programs & Practices in the Public Schools* (1986).

TABLE 1. Number of Standardized Tests Administered in the Public Schools To Fulfill State Mandates, By State (1986-1987 School Year).

STATE	NUMBER OF TESTS	SCHOOL ENROLLMENT	STATE-MANDATED TESTS PER STUDENT
Alabama	720,000	733,735	0.98
Alaska	0	107,973	0.00
Arizona	480,000	534,538	0.90
Arkansas	179,000	437,438	0.41
California	1,420,000	4,377,989	0.32
Colorado	208,800	558,415	0.37
Connecticut	99,000	468,847	0.21
Delaware	60,000	94,410	0.64
District of Columbia	119,000	85,612	1.39
Florida	437,352	1,607,320	0.27
Georgia	1,020,000	1,096,425	0.93
Hawaii	75,000	164,640	0.46
Idaho	30,000	208,391	0.14
Illinois	418,000	1,825,185	0.23
Indiana	211,000	966,780	0.22
Iowa	0	481,286	0.00
Kansas	135,000	416,091	0.32
Kentucky	650,000	642,778	1.01
Louisiana	390,000	795,188	0.49
Maine	48,000	211,752	0.23
Maryland	168,000	675,747	0.25
Massachusetts	410,000	833,918	0.49
Michigan	326,285	1,681,880	0.19
Minnesota	0	711,134	0.00
Mississippi	240,000	498,639	0.48
Missouri	664,000	800,606	0.83
Montana	0	153,327	0.00
Nebraska	20,508	267,139	0.08
Nevada	55,000	161,239	0.34
New Hampshire	30,000	163,717	0.18
New Jersey	630,000	1,107,467	0.57
New Mexico	80,000	281,943	0.28
New York	1,691,000	2,607,719	0.65
North Carolina	1,085,000	1,085,248	1.00
North Dakota	0	118,703	0.00
Ohio	0	1,793,508	0.00
Oklahoma	234,000	593,183	0.39
Oregon	15,000	449,307	0.03
Pennsylvania	538,212	1,674,161	0.32
Rhode Island	40,000	134,126	0.30
South Carolina	450,000	611,629	0.74
South Dakota	50,000	125,458	0.40
Tennessee	500,000	818,073	0.61
Texas	1,500,000	3,209,515	0.47
Utah	258,000	415,994	0.62
Vermont	0	92,112	0.00
Virginia	377,000	975,135	0.39
Washington	182,000	761,428	0.24
West Virginia	200,000	351,837	0.57
Wisconsin	309,000	767,819	0.40
Wyoming	0	100,955	0.00
UNITED STATES	16,753,157	39,837,459	0.42

Table 2. Standardized Tests Administered in the Public Schools To Fulfill Local Mandates (1986-1987 School Year).

SCHOOL DISTRICT	NUMBER OF TESTS	SCHOOL ENROLLMENT	DISTRICT-MANDATED TESTS PER STUDENT
CALIFORNIA			
Los Angeles	949,899	540,903	1.76
San Juan Unified	33,000	44,186	0.75
San Francisco	21,000	58,378	0.36
San Diego City Unified	34,040	110,631	0.31
COLORADO			
Jefferson County	52,500	76,000	0.69
CONNECTICUT			
Greenwich	5,720	6,772	0.84
FLORIDA			
Hillsborough County (Tampa)	98,000	115,323	0.85
Pinellas County (Clearwater)	112,000	85,339	1.31
Duvall County (Jacksonville)	190,000	99,512	1.91
Dade County (Miami)	230,000	250,000	0.92
Broward Cty. (Ft. Lauderdale)	336,000	129,478	2.60
GEORGIA			
Fulton County (Atlanta)	13,000	35,523	0.37
DeKalb County (Decatur)	15,000	66,000	0.23
ILLINOIS			
Chicago	468,544	435,000	1.08
INDIANA			
Indianapolis	66,600	50,600	1.32
IOWA			
Des Moines Independent	44,500	30,000	1.48
KANSAS			
Wichita	65,729	44,729	1.47
KENTUCKY			
Jefferson County (Louisville)	47,000	95,020	0.49
LOUISIANA			
Orleans Parish (New Orleans)	84,000	84,000	1.00
MARYLAND			
Baltimore City	275,800	120,000	2.30
Prince Georges County	175,000	103,000	1.70
MASSACHUSETTS			
Boston	60,000	55,000	1.09
Brookline	3,500	5,400	0.65
MICHIGAN			
Detroit City	269,000	200,000	1.35

SCHOOL DISTRICT	NUMBER OF TESTS	SCHOOL ENROLLMENT	DISTRICT-MANDATED TESTS PER STUDENT
MINNESOTA			
Minneapolis	39,712	32,274	1.23
MISSOURI			
St. Louis	119,327	48,800	2.45
MONTANA			
Missoula	6,990	5,640	1.24
NEVADA			
Las Vegas	87,684	95,000	0.93
NEW HAMPSHIRE			
Concord	7,296	5,000	1.46
NEW JERSEY			
Newark	603,000	67,000	9.00
NEW MEXICO			
Albuquerque	133,320	80,000	1.67
NEW YORK			
New York City	1,133,000	924,123	1.23
Rochester	35,000	34,696	1.01
Buffalo City	63,000	44,707	1.41
NORTH CAROLINA			
Mecklenburg Cty. (Charlotte)	18,000	72,162	0.25
Wake County (Raleigh)	40,500	58,213	0.70
NORTH DAKOTA			
Fargo	6,000	9,200	0.65
OHIO			
Cincinnati	123,800	53,000	2.34
Akron	20,520	34,000	0.60
Cleveland	219,000	75,000	2.92
OKLAHOMA			
Oklahoma City	31,300	40,000	0.78
OREGON			
Portland	30,000	52,000	0.58
PENNSYLVANIA			
Philadelphia	400,000	200,000	2.00
SOUTH CAROLINA			
Greenville County	37,000	53,000	0.70
TENNESSEE			
Memphis City	188,200	106,000	1.76

SCHOOL DISTRICT	NUMBER OF TESTS	SCHOOL ENROLLMENT	DISTRICT-MANDATED TESTS PER STUDENT
TEXAS			
Dallas Independent	217,584	127,584	1.71
Houston Independent	250,000	193,702	1.29
UTAH			
Salt Lake City	56,000	72,000	0.78
VERMONT			
Burlington	2,212	3,800	0.58
VIRGINIA			
Fairfax County	10,000	124,631	0.08
Virginia Beach City	0	54,870	0
Prince William County	105,000	57,213	1.84
WASHINGTON			
Seattle	41,500	44,000	0.94
WEST VIRGINIA			
Kanawha County (Charleston)	20,300	37,399	0.54
WISCONSIN			
Milwaukee	64,500	96,387	0.67
WYOMING			
Laramie County (Cheyenne)	11,000	13,000	0.84

TABLE 3. COMPUTATION OF ESTIMATE OF STANDARDIZED TESTS ADMINISTERED BY PUBLIC SCHOOLS TO FULFILL STATE OR LOCAL TESTING MANDATES (1986-87 SCHOOL YEAR).

	SURVEYED			ESTIMATED		TOTAL	
TYPE OF SYSTEM	STUDENTS	TESTS	RATE	STUDENTS	TESTS	STUDENTS	TESTS
LARGE (OVER 100,000)	3,680,375	5,035,067	1.37	---	---	3,680,375	5,035,067
MEDIUM (25,000 TO 100,000)	2,026,008	2,790,792	1.11	4,674,000	5,188,000	6,700,008	7,978,792
SMALL (LESS THAN 25,000)	48,812	42,718	0.88	29,408,000	25,879,000	29,456,812	25,921,718

TESTS ADMINISTERED TO FULFILL —	TESTS IN SURVEY	ESTIMATED TESTS	TOTAL TESTS
STATE MANDATES	16,753,157	—	16,753,157
LOCAL MANDATES	7,868,577	31,067,000	38,935,577
ALL MANDATES	24,621,734	31,067,000	55,688,734

NOTE: The computation of the testing rate in medium-sized school systems excludes the unusually high testing rate of Newark, New Jersey (which is three times higher than the next highest system.) The estimated number of tests administered in medium and small school systems was computed using the actual rate administered for medium and small systems based on survey results.

Appendix B
What Is Authentic Evaluation?

Authentic evaluation of educational achievement directly measures actual performance in the subject area. Standardized multiple-choice tests, on the other hand, measure test-taking skills directly, and everything else either indirectly or not at all.

Also called "performance," "appropriate," "alternative," or "direct" assessments, authentic evaluation includes a wide variety of techniques, such as written products, solutions to problems, experiments, exhibitions, performances, portfolios of work and teacher observations, checklists and inventories, and cooperative group projects. They may be the evaluation of regular classroom activity or take the form of tests or special projects.

Authentic evaluations indicate what we value by directing instruction toward what we want the student to know and be able to do. They are appropriate to the student's age and level of learning and the subject being measured, and are useful to both teachers and students.

All forms of authentic assessment can be summarized numerically or put on a scale. Therefore, individual results can be combined to provide a variety of information about aggregate performance at the classroom, school, district, state and national levels. Thus, state and federal requirements for comparable quantitative data can be met.

Authentic assessment was developed in the arts and in apprenticeship systems, where assessment has always been based on performance. It is impossible to imagine evaluating a musician's ability without hearing her or him sing or play an instrument, or judging a woodworker's craft without seeing the table or cabinet. It is also impossible to help a student improve as a woodworker or musician unless the instructor observes the student in the process of working on something real, provides feedback, monitors the student's use of the feedback, and adjusts instruction and evaluation accordingly. Authentic assessment extends this principle of evaluating real work to all areas of the curriculum.

The most widely used form of authentic assessment in education today is in writing. For example, twenty-eight states, the National Assessment of Educational Progress (NAEP), and many other na-

tions ask students to write on assigned topics. The essays and stories are graded by teams of readers (usually teachers) who assign grades according to standard guidelines. The readers are trained and retrained throughout the process to maintain reliable standards, a process that produces a high degree of agreement among judges. As with all the examples, this methodology can be used to evaluate classroom work that has been collected in a portfolio; it only has to be adjusted for subject area and student age.

Similar procedures are now being followed with open-ended mathematics questions. These ask students to write their own response (not just the answer) to a problem. There is no single way to find a "right answer" because the question is designed to see how a student thinks through a problem, thereby indicating her or his ability to use math. The answers are scored by groups of teacher-readers, again following a standard grading procedure. Two-sevenths of the NAEP math questions will be open-ended in 1990.

Performance assessments in science ask students to plan or perform experiments or use scientific apparatus, as is done in New York State. Science assessments can be graded by observation (where a teacher or other observer uses a checklist) or by scoring the students' written answers to the questions. These assessments can be developed to indicate understanding of basic scientific concepts and methods.

History/social studies assessments frequently require group projects, such as preparing a history of the neighborhood or discovering how a group of people changed a law or policy, tasks which allow students to demonstrate that they grasp important concepts about history and about democratic processes. Foreign language assessments ask students to use the language in a real-life situation, orally and in print.

For young students, reading is best evaluated by having a student read aloud from material of varying levels of difficulty, while keeping a record of "miscues" that reveal the reader's strengths and weaknesses and the strategies used to solve problems. The reading can be taped and reviewed by teacher and student for further analysis and to monitor progress. For older and younger students, the material can be discussed to evaluate comprehension and critical thinking. A writing assignment responding to the ideas of the reading passage can reveal the student's proficiency and thinking in both reading and writing.

All these assessments can be designed to closely follow the curriculum. They provide continuous, qualitative data that can be used by teachers to help instruction. They can be used by students, who learn to assume responsibility for their portfolios and records and thereby engage in regular self-analysis of their work and progress. They provide a direct measure of achievement and therefore are worth the time spent preparing for and doing them. They also encourage an intelligent, rich curriculum rather than the dumbed-down, narrow curriculum fostered by teaching to and coaching for multiple-choice tests.

Teachers can and should be assisted in the evaluation process by community groups, parents, administrators, and university faculty. Outside participation can ensure that racial or cultural bias does not distort the assessment process. For example, a team can examine student portfolios and then compare their evaluations with those of the teacher. These teams should also be helpful in strengthening the evaluation capabilities of teachers by providing feedback.

Authentic evaluation will provide far more information than any multiple-choice test possibly could. The costs of teacher involvement in designing, administering, and scoring new assessments can be counted as part of professional and curriculum development, since no other activity involves teachers more deeply in thinking about their teaching, its objectives, methods and results. Schools and communities will see that authentic assessments are promoting the thinking curriculum everyone wants for our children, and thereby providing genuine accountability.

PRE-SCHOOL AND K-12 TESTING: ANNOTATED BIBLIOGRAPHY

1. MATERIALS FROM FAIRTEST

FairTest — *FairTest Examiner.*

Quarterly newsletter surveys developments in testing and testing reform, including pre-school, grade school, IQ and related tests. (Available from FairTest, $15/yr individuals; $25/yr institutions.)

Neill, D.M. and Medina, N.J. — **"Standardized Testing: Harmful to Educational Health."** ***Phi Delta Kappan*** **(May 1989) pp. 688-697.**

Similar to *Fallout from Testing Explosion* in content, though shorter, no tables on extent of test use and no annotated bibliography. Included in a special section in this *Kappan* on testing and alternatives to testing (see 5. Authentic Evaluation, below).

FairTest has the following fact sheets available (send SASE with request for fact sheet): "What's Wrong with Standardized Tests," "How Standardized Testing Harms Education," "What Is Authentic Assessment?" "Testing Reform in Grades 1 and 2, North Carolina."

For other FairTest publications, see the order form at the back of this report.

2. GENERAL DISCUSSIONS OF TESTING

(Many of these articles are also relevant to the subsections that follow.)

Airasian, P. W. — **"Measurement Driven Instruction,"** ***Educational Measurement*** **(Winter 1988) pp. 6-11.**

Under some conditions, measurement driven instruction may corrupt the measurement process.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Standards for Educational and Psychological Testing **(Washington, DC: APA, 1985).**

The testing profession's latest revision of its guidelines for the proper construction and use of standardized tests. This work is a useful tool in reform efforts because much test construction and use fails to meet even these weak guidelines.

Bastian, A., *et.al.*

Choosing Equality: The Case for Democratic Schooling **(Philadelphia: Temple University Press, 1986).**

Concludes that democratizing schools to meet the needs of all students ought to be the focus of school reform efforts. Standardized tests are shown to be one of the barriers to equal educational opportunity and educational quality.

Congressional Budget Office

Trends in Educational Achievement **(Washington, D.C.: U.S. Government Printing Office, April 1986).**

The first of two CBO reports on educational achievement. Includes general discussion of the ways in which standardized test results are used to measure educational achievement and the problems with such uses. Warns against overreliance on test results in evaluating trends in educational achievement.

Fiske, E.

"America's Test Mania." ***New York Times*** **(April 10, 1988) Section 12, pp. 16 - 20.**

Contains general overview of the growth of standardized testing in America's public schools and the resulting problems. Discusses various uses made of standardized tests, efforts by schools to improve test scores, and criticisms and limitations of standardized tests. It also suggests that standardized tests will undermine the goals of the current school reform movement.

Gould, S.J.

The Mismeasure of Man **(New York: Norton, 1981).**

Criticizes the idea that there is a single unitary thing that can be called "intelligence" and the tendency to reduce information to a single number. Provides a history of "mismeasures," including "IQ" tests. Highly readable, valuable critique.

Haney, W.

"Test Reasoning and Reasoning about Testing." ***Review of Educational Research*** **(Winter 1984) p. 628.**

Provides a detailed history of the use of standardized tests in America (divided into three eras: pre-WWI; WWI to 1950; and 1950

to present). In particular, chronicles the recent growth in test use and test criticism. Includes brief discussion of the intensity of use of standardized tests in the public schools and a longer discussion of the types of uses. Concludes by suggesting avenues for further research and development in testing.

Haney, W. & G. Madaus

"Effects of Standardized Testing and the Future of the National Assessment of Educational Progress." (Working Paper for the NAEP Study Group, 1986. ERIC Document ED-279-680.)

Documents the increasing attention paid to testing compared with curriculum in educational literature. Discusses four broad issues which determine the impact of testing on education and lists seven major problems regarding the impact of tests on individuals and schooling. Authors suggest efforts that can minimize the negative impact and maximize the possible benefits of testing. Two sections specifically discuss NAEP and its future.

Madaus, G. F.

"The Influence of Testing on the Curriculum." ***87th Yearbook of the National Society for the Study of Education, Part I: Critical Issues in the Curriculum*** **(1988) pp. 83-121.**

An excellent, comprehensive look at the effects of testing on the curriculum. After describing the various types of tests and testing programs, Madaus poses seven general principles describing the impact of testing on the curriculum. He argues that tests can become the "ferocious master" instead of being the "compliant servant," and details the evidence of their effects on programs, teachers and students.

Madaus, G. & D. Pullin

"Questions to Ask When Evaluating a High-Stakes Testing Program." ***NCAS Backgrounder*** **(Boston: National Coalition of Advocates for Students, June 1987).**

Presented in question-and-answer format; focuses on use of standardized tests in "high stakes" educational situations (i.e. where significant sanctions or rewards are associated with the test). Lists several "high stakes" uses of tests, discusses ways in which tests are designed or selected. Examines potential structure of testing programs, possible conclusions that will be drawn from their results, and their likely impact on students and schools. (NCAS, 100 Boylston St., Boston, MA 02116; free.)

National Elementary Principal

(March/April 1975 and July/August 1975).

The first issue, "IQ: The Myth of Measurability," contains a variety of criticisms of IQ tests. The second, "The Scoring of Children: Standardized Testing in America," examines the testing industry, test construction, and achievement tests.

Negro Educational Review

Special Issue: "Testing African American Students" (April - July 1987).

Book-length issue contains numerous articles on testing, including discussions of: psychometric, language and cultural biases against blacks (particularly working class blacks) and other minorities; IQ testing; and alternatives to standardized tests. Several specific articles are noted below. (Available from *NER*, Box 2895, Jacksonville, FL 32203; $20). The July-October 1977 issue of this journal, "Testing Black Students," also includes many excellent articles.

Ninth Mental Measurement Yearbook

(Lincoln, Nebraska: Buros Institute of Mental Measurement, 1985).

Contains one or more reviews plus bibliography for each of hundreds of tests; older tests may also be in previous editions of the *Yearbook*. The primary source for test reviews.

Phi Delta Kappan

"What is the Proper Role of Testing?" (Special Section). (May 1985) pp. 599 - 639.

Nine articles focus on the use, problems and impact of standardized testing in the public schools. Topics covered include: misuse of SAT scores in assessing the quality of American education; popularity of standardized tests; impact of testing on pedagogy and instruction; problems with teacher testing; impact of testing on educational equity; and alternatives to multiple-choice writing tests. One article describes the use of standardized tests to "drive the curriculum" in three states and one school district.

Pipho, C.

"Tracking the Reforms: Part 5 - Testing." ***Education Week*** **(May 22, 1985) p. 19.**

One in a series of articles discussing the state education reform efforts of 1983 - 1985. Lists state uses of standardized tests.

Rethinking Schools

(Jan/Feb 1989).

"Focus on Testing" section includes critiques of testing and its effects, and a discussion of high school alternatives. (P.O. Box 93371, Milwaukee, WI 53202; $2.00.)

3. SPECIFIC PROBLEMS WITH STANDARDIZED TESTS

A. TESTING YOUNG CHILDREN

Bredekamp, S. and L. Shepard

"How Best to Protect Children from Inappropriate School Expectations, Practices and Policies." ***Young Children*** **(March 1989) pp. 14-24.**

Contains excellent critique of the misuses of standardized testing on young children as well as discussion of appropriate use. Adds to NAEYC publications noted below. Includes helpful bibliography.

Kamii, C., ed.

Achievement Testing in the Early Grades: The Games Grown-Ups Play **(Washington: National Association for the Education of Young Children, 1990).**

This book summarizes the problems caused by reliance on multiple-choice achievement testing, has chapters on specific problems for systems, principals, teachers and others, and examines the problems testing causes for teaching literacy and math. Appropriate assessment in math and literacy is also presented.

Martin, A.

"Screening, Early Intervention, and Remediation: Obscuring Children's Potential," ***Harvard Educational Review*** **(Nov. 1988) pp. 488-501.**

Finds disability testing and special needs intervention frequently produce an overemphasis on children's weaknesses and promote a view of children as having deficits to be corrected, rather than having individual differences and strengths on which to build. The same factors disempower classroom teachers.

Meisels, S.J.

"Uses and Abuses of Developmental Screening and School Readiness Testing." ***Young Children*** **(Jan. 1987) pp. 4-6, 68-73.**

Looks at different tests and their limitations, urges caution in their use. Specifically examines the Gesell instruments. The issue contains a response from Gesell and a rebuttal by Meisels. A major source for the NAEYC position. For a specific discussion of screening tests and a review of a number of tests (many of which lack adequate validity), see Meisels, S.J. *Developmental Screening in Early Childhood: A Guide* (NAEYC, Third Edition, 1989).

National Association for the Education of Young Children

"NAEYC Position Statement on Standardized Testing of Young Children 3 Through 8 Years of Age." ***Young Children*** **(March 1988) pp. 42-47.**

Reviews appropriate and inappropriate test use. Urges caution in the use of tests. Basis for pamphlet (below). See also, "NAEYC Position Statement on Developmentally Appropriate Practice in the Primary Grades, Serving 5- Through 8-Year-Olds," *Young Children* (Jan. 1988) p. 64-84, which includes section on testing and has a comprehensive bibliography.

National Association for the Education of Young Children

Testing of Young Children: Concerns and Cautions **(Washington: NAEYC, 1988).**

Pamphlet discusses the potentially harmful impact on young children of standardized testing. Describes proper uses of standardized tests and suggests how schools can "help ensure that all children get off to a sound start in kindergarten, first, and second grade." (Available from NAEYC, 1834 Connecticut Avenue, N.W., Washington, D.C. 20009; $.50 each or 100 for $10.)

Shepard, L.A. and M.L. Smith, eds.

Flunking Grades **(London: Falmer, 1989).**

Chapters 4 and 5 explore the harm of retaining young children in grade or placing them in "transitional programs," often done through the use of test results. See also by the same authors, "Flunking Kindergarten," *American Educator* (Summer 1988) pp. 34-38, and "What Doesn't Work: Explaining Policies of Retention in the Early Grades," *Phi Delta Kappan* (October 1987).

B. TEST BIAS

First, J.M. and J. Willshire Carrera

New Voices: Immigrant Students in U.S. Public Schools **(Boston: National Coalition of Advocates for Students, 1988).**

Includes discussion of the harmful effects of standardized testing on new immigrant students. (NCAS, 100 Boylston St., Boston, MA 02116; $11.95).

Hilliard, A.

"Ideology of I.Q." ***Negro Educational Review*** **(April-July 1987) pp. 136-145 (see** ***Negro Ed. Review*****, Section 2., above).**

A good discussion of the basis of I.Q. testing.

Hoover, M.R., R.L. Politzer & O. Taylor

"Bias in Reading Tests for Black Language Speakers: A Sociolinguistic Perspective." ***Negro Educational Review*** **(April-July 1987) pp. 81 - 98 (see** ***Negro Ed. Review,*** **Section 2., above).**

Details language-related bias in standardized tests against speakers of non-standard English, including phonological (sound), syntactical (structural), and lexical (word choice and vocabulary) biases. Consequences of bias include school program misplacement and tracking resulting in inadequate education for students who are not white middle- to upper-class. Eliminating these biases is important for reducing educational and societal biases against working class and minority children.

Levidow, L.

"'Ability' Labeling as Racism." In D. Gill and L. Levidow, ***Anti-Racist Science Teaching*** **(London: Free Association Books, 1987) pp. 231-267.**

Summarizes a number of problems with I.Q. tests, including bias in the tests, and argues they are harmful social constructs. Urges caution in using other methods of assessment as they can be as damaging as tests.

Sosa, A.S.

The Impact of Testing on Hispanics **(Berkeley: National Commission on Testing and Public Policy, 1988).**

Wide-ranging document contains abstracts of two major papers plus testimony to the Commission. One paper discusses secondary education; seven witnesses address elementary and secondary education.

Taylor, O. & D.L. Lee

"Standardized Tests and African Americans: Communication and Language Issues." ***Negro Educational Review*** **(April-July 1987) pp. 67-80 (See** ***Negro Ed. Review,*** **above).**

Contains detailed discussion of sources and kinds of cultural and language bias in standardized tests. These biases cause African-Americans (particularly working-class blacks) and minorities to be invalidly assessed: "At times. . . the results fail to accurately represent actual abilities." In conclusion, ". . . the very assumptions and paradigms upon which most standardized tests are based need to be revised."

Willie, C.V.

"The Problem of Standardized Testing in a Free and Pluralistic Society." ***Phi Delta Kappan*** **(May 1985) pp. 626-628.**

Discusses disproportionate impact of test use on minority students. Argues that tests ignore the diversity among various American ethnic groups and implicitly undercut the value and legitimacy of this diversity.

C. TEST VALIDITY & RELIABILITY

Cannell, J.J. — ***Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States Are Above the National Average*** **(Friends for Education: Daniels, W.Va., 1987).**

Describes the results of a nationwide survey on achievement test scores: no state average score at the elementary level was below the national norm on any of the six most popular achievement tests. Concludes that this results from improper norming of the tests and teaching to the tests. (For confirming studies, see Notes 39 and 40 at end of main body of *Fallout*, and Koretz, below).

Congressional Budget Office — ***Educational Achievement: Explanations and Implications of Recent Trends*** **(Washington, D.C.: U.S. Government Printing Office, August 1987).**

Second of two CBO reports on educational achievement, includes a general discussion of the problems and limitations of standardized tests, particularly the problems around validity and reliability.

Johnston, P. — **"Constructive Evaluation and the Improvement of Teaching and Learning."** ***Teachers College Record*** **(Summer 1989) pp. 509 - 528.**

Criticizes the search for "objectivity" as both futile and educationally destructive. Points out that the core of validity is construct validity, which is a "socially negotiated variable within a changing pluralistic society." Concludes that "the overwhelming concern for objective, valid measurement as a means to improve teaching and learning is not very helpful." Discusses the need for a "different view of science" and for improving the quality of teachers. Concludes with various concrete examples of teaching and evaluating. (For a different view, calling for expanding validity studies to include teacher and learner use of tests, see C.K. Tittle, "Validity: Whose Construction Is it in the Teaching and Learning Context," *Educational Measurement*, Spring 1989, pp. 5-13, 34.) The two articles also provide excellent bibliographies of current validity discussions.

Koretz, D. — **"Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction?"** ***American Educator*** **(Summer 1988) pp. 8-15, 46-52.**

Analyzes and substantially confirms Cannell's report (above) and essentially answers "Yes" to the questions it poses in the title. Calls for changing tests and re-focusing on the broader goals of education, not just testing.

Messick, S. **"Meaning and Values in Test Validation,"** ***Educational Researcher*** **(March 1989) pp. 5-11.**

The most recent in a series on validity by Messick. Argues for construct validity as the essence of validity, then expands the definition of construct validity to include value implications and social consequences; e.g., if use of a test in a particular situation has harmful consequences, the use of the instrument could lack construct validity if the harm is caused in part by the nature of the test itself. See also by Messick: "The Once and Future Issues of Validity," in Howard Wainer and Henry I. Braun, eds., *Test Validity* (Hillsdale, NJ: Lawrence Erlbaum, 1988); and his lengthy, technical article in Robert Linn, ed., *Educational Measurement* (New York: MacMillan, 1989).

D. TEST ADMINISTRATION

Fuchs, D. & L.S. Fuchs **"Test Procedure Bias: A Meta-Analysis of Examiner Familiarity Effects."** ***Review of Educational Research*** **(Summer 1986) pp. 243-262.**

Authors analyze data from 22 controlled studies involving 1489 subjects and discover that familiarity with the test examiner had different impacts on test-takers. Use of unfamiliar examiners reduces test scores for low-income and black students, but does not affect the scores of upper-income students.

Wodtke, K., *et al.* **"Social Context Effects in Early School Testing: An Observational Study of the Testing Process." (Paper presented to the American Educational Research Association Annual Convention, 1985).**

Study examines the administration of standardized tests in eight kindergartens, finding that "testing practices in five of the eight kindergartens were so nonstandardized as to render their test scores incomparable and quite possibly unreliable as well." Concludes that using results from such administrations is likely to lead to unsound educational decisions.

4. THE IMPACT OF TESTING ON EDUCATION

Section 2 (above) also includes items relevant to this topic.

Madaus, G. **"Test Scores as Administrative Mechanisms in Educational Policy." Phi Delta Kappan (May 1985) pp. 611-618.**

Provides very brief history of testing in American public schools, suggests reasons for the increased use of tests in the schools. Describes

various problems with test use including: loss of authority for teachers' professional judgments; loss of local control over education; narrowing the educational curriculum; teaching to the test; disproportionate impact on disadvantaged students; and elevation of coaching over teaching.

A. IMPACT ON CURRICULUM

Allington, R.L.

"The Reading Instruction Provided Readers of Differing Reading Abilities." ***Elementary School Journal*** **(Vol. 83, #5) pp. 548-559.**

Finds that good readers are taught differently from poor readers. Poor readers experience lots of oral reading (with regular interruptions), decoding and word emphasis, with absolute test score gains as the measure. Good readers experience silent reading and comprehension discussions, both of which correlate with quicker and stronger comprehension improvement (decoding emphasis leads to increases in decoding on tests). Recommends more silent reading, summary and discussion for poor readers, plus more research, as this report is often tentative. See also by Allington: "Poor Readers Don't Get to Read Much in Reading Groups," *Language Arts* (Nov.-Dec. 1980), and "Shattered Hopes: Why Two Federal Reading Programs Have Failed to Correct Reading Failure," *Learning* (July-Aug. 1987).

Bussis, A. M.

"'Burn It at the Casket': Research, Reading Instruction, and Children's Learning of the First R." ***Phi Delta Kappan*** **(December 1982) pp. 237-241.**

"Skills" instruction displaces real reading in classrooms despite lack of evidence that "*any* identifiable group of subskills was essential to reading," (according to IRA study in 1973 and NIE study in 1975). "Only after educators began to realize that many children could master decoding skills and still fail to read effectively did 'the focus of research begin to expand'." Presents constructionist psychological view, distinguishes information from meaning. The skill of reading is complex, but not an assemblage of subskills. Different children learn differently. Concludes with teaching suggestions focused on real reading and increasingly observant teachers.

Bussis, A.M. and E.A. Chittendon.

"What the Reading Tests Neglect." ***Language Arts*** **(March 1987) pp. 302-308.**

Excellent summary of research indicating "little correspondence between contemporary theories of the reading process and assumptions implicit in the tests."

Chittendon, E.A.

"Styles, Reading Strategies and Test Performance: A Follow-Up Study of Beginning Readers," in R.O. Freedle and R.P. Duran, eds., *Cognitive and Linguistic Analyses of Test Performance* (Norwood, N.J.: Ablex, 1987).

Based on a six-year study of how children learn to read, compares test item types to children's differing strategies in learning to read, finds much incompatibility. Connects this research to more general incompatibility between tests and current cognitive, developmental and learning theory.

Dorr-Bremme, D.W. and J.L. Herman

***Assessing Student Achievement: A Profile of Classroom Practices* (Los Angeles: Center for the Study of Evaluation, 1986).**

Reports on a nation-wide survey of teachers and principals in 114 school districts. Seeks to identify the amount of time devoted to testing, uses of test results, teachers' and principals' perceptions about testing, and issues of equity as a result of standardized test use. Concerns about impact of standardized test use include the fact that most teachers pay more attention to standardized test results for low SES students than for high SES students. Tests also can reduce time spent on other educational goals and narrow curriculum. Calls for more "rational" relationship between teacher-designed tests, external texts and the curriculum, but emphasizes that the curriculum must drive the tests and not vice-versa.

Frederiksen, N.

"The Real Test Bias: Influences of Testing on Teaching and Learning." *American Psychologist* (March 1984) pp. 193-202.

From abstract: "There is evidence that tests do influence teacher and student performance and that multiple-choice tests tend not to measure the more complex cognitive abilities." Describes several experiments showing that open-ended (free response, multiple-level) problem-solving tests measure different, higher and more complex, cognitive abilities than do multiple-choice tests, even multiple-choice tests directly derived from the open-ended tests. Most of the problems in the world are ill-structured, unlike multiple-choice problems which dominate tests used for accountability. "We need a much broader conception of what a test is if we are to use test information in improving educational outcomes."

Goodman, K.S. *et al.*

***Report Card on Basal Readers* (Katonah, NY: Richard C. Owen, 1988).**

Prepared for the National Council of Teachers of English. Critical analysis of the history and content of basal readers and how they hinder the teaching of reading. Includes discussion of tests contained in basals and the relationship between basals and standardized tests.

Kamii, C. — *Young Children Continue to Reinvent Arithmetic, 2nd Grade* (New York: Teachers College Press, forthcoming).

Chapter 10, on evaluation, indicates that standardized tests cannot reveal whether students understand math concepts and reasoning. Teaching to the tests makes it less likely that students develop quantitative reasoning concepts and skills.

LeMahieu, P. G. and R. C. Wallace — "Up Against the Wall: Psychometrics Meets Praxis." *Educational Measurement* (Spring 1986) pp. 12-16.

"It is untenable to agree that achievement is the product, and that test scores are its measure, and then assert, 'Please don't pay *too* much attention to the scores.'" Discusses test use in Pittsburgh, where the focus is diagnostic testing that is relevant, timely (quick turn-around, frequent), short, used by teachers to inform instruction. Pittsburgh uses regular achievement tests for evaluative, not diagnosic purposes, and only uses samples of students. Testing is still mostly multiple-choice.

McClellan, M.C. — "Testing and Reform." *Phi Delta Kappan* (June 1988) pp. 768-771.

Short discussion of the effects of testing includes the argument, with references, that test-based methods used to teach basic skills (rote, drill, retention) are counter productive to teaching and learning higher order thinking skills.

McNeil, L.M. — "Contradictions of Reform." *Phi Delta Kappan* (March 1988) pp. 478-485.

Part of a series by McNeil, "Contradictions of Control." Explores the destructive effects of mandated testing on an exceptionally good school program.

Meier, D. — "Why Reading Tests Don't Test Reading." *Dissent* (Winter 1982-83).

Shows that reading instruction aimed at increasing standardized test scores hinders learning to read, since higher test scores do not necessarily indicate improved reading ability. Argues that reading and other tests are biased against minority and low-income youth. Tests cannot measure the utility or effect of school reforms: "... testing not only fails to be helpful but sabotages good education."

National Academy of Education

The Nation's Report Card: Improving the Assessment of Student Achievement, Review of the Alexander/James Study Group Report **(Cambridge, MA: National Academy of Education, 1987).**

Although this review discusses the recommendations of the Alexander/James Study Group to dramatically expand the National Assessment of Educational Progress (NAEP), it also provides some general discussions of the problems of testing. In particular, it notes how testing can narrow the curriculum, threaten educational equity, and undermine educational goals that are not easily quantifiable.

Resnick, L.B. and D.P. Resnick

"Assessing the Thinking Curriculum: New Tools for Educational Reform," in B.R. Gifford and M.C. O'Connor, eds. ***Future Assessments: Changing Views of Aptitude, Achievement, and Instruction*** **(Boston: Kluwer Academic Publishers, 1989).**

Shows how tests are based on outmoded behaviorist and associationist psychological theories from late 19th century in which knowledge is reduced to isolated bits and learning is defined as passively absorbing the bits; current theory indicates knowledge is whole and contextual, learning is active, contextual, and constructed in people's minds. Current tests are incompatible with needed educational reforms. New forms of assessment must be based on accurate learning theory and help school reform.

Salganik, L.H.

"Why Testing Reforms Are So Popular and How They Are Changing Education." ***Phi Delta Kappan*** **(May 1985) pp. 607-610.**

Links the growing use of standardized tests in the schools with a loss of public confidence in teachers. Because increased reliance on testing undermines the authority of teachers' judgments, a cycle of declining professional authority and declining public confidence is created. Policy issues, such as educational equity, the goals of schooling and control over school decisions, are masked by the emphasis on technical questions regarding testing.

Smith, F.

Insult to Intelligence: The Bureaucratic Invasion of Our Classrooms **(New York: Arbor House, 1986).**

Excellent, readable book on learning, reading, cognitive theory and how testing and much reading instruction are insulting to the intelligence of children (and teachers), who then turn off to reading and schooling. Also has much on good methods of teaching and how to reform education.

Suhor, C.

"Objective Tests and Writing Samples: How Do They Affect Instruction in Composition?" ***Phi Delta Kappan*** **(May 1985) pp. 635-639.**

Criticizes the use of multiple-choice standardized writing tests

because they undercut "real" writing instruction. Urges use of computerized "writing sample" tests as an alternative.

Tyson-Bernstein, H. ***A Conspiracy of Good Intentions: America's Textbook Fiasco*** **(Washington, D.C.: Council for Basic Education, 1988).**

Discusses the quality of textbooks in the public schools. Concludes that the increasing emphasis on testing has been a major contributor to the declining quality of American textbooks. Indicts the current "curriculum alignment" movement which has affected school districts in 22 states. Encourages the use of a more diverse curriculum and set of student assessment mechanisms. (CBE, 725 15th St., N.W., Washington, D.C. 20005; $13.)

B. IMPACT ON STUDENT PROGRESS AND ACHIEVEMENT

Chunn, E.W. **"Sorting Black Students for Success and Failure: The Inequity of Ability Grouping and Tracking."** ***Urban League Review*** **(Vol. 11, No. 1&2, 1987) pp. 93-106.**

Summarizes the harmful effects of tracking on black students. For the effects of tracking on segregation, see also J.L. Epstein, "After the Bus Arrives: Resegregation in Desegregated Schools," *Journal of Social Issues* (Vol. 41, No. 3, 1985) pp. 23-43.

Oakes, J. ***Keeping Track: How Schools Structure Inequality*** **(New Haven: Yale University Press, 1985).**

The key work on the harmful effects of tracking in the public schools, often done on the basis of test scores. Tracking is most harmful to low-income and minority-group children because of their isolation from other children and the watered-down curriculum they receive. See also Oakes' article, "Keeping Track," *Phi Delta Kappan* (Sept. & Oct., 1986).

Pullin, D. **"Educational Testing: Impact on Children At Risk."** ***NCAS Backgrounder*** **(Boston: National Coalition of Advocates for Students, December 1985).**

This report discusses: increased use of standardized tests in the public schools; their use as barriers to educational opportunity for at-risk children; misclassification of minority students; impact of tests on handicapped students; and the general measurement limitations of tests. (NCAS, 100 Boylston St., Boston, MA 02116; free.)

Raudenbush, S.

"Magnitude of Teacher Expectancy Effects on Pupil IQ as a Function of the Credibility of Explanatory Induction — A Syntheses of Findings from 18 Experiments." ***Journal of Educational Psychology*** **(Vol. 76, No. 1, 1984) pp. 85-97.**

Reanalyzes the "Pygmalion Effect," originally identified and described by Robert Rosenthal and Lenore Jackson in 1965 in *Pygmalion in the Classroom*. Discovers "Effect" prevalent among students entering seventh grade, not those in grades three to six. "Effect" most likely to occur when teachers know little about their students beyond the statistical information they are provided. Where that information (including test scores) under- or overestimates students' abilities, teacher behavior affects student achievement (regardless of previous student achievement levels).

Shepard, L. A. and M.L. Smith, eds.

Flunking Grades **(cited above, 3A Young Children).**

While Chs. 4 & 5 discuss the effects of retention on young children, the book as a whole surveys the extent of retention, why it happens—including test-based retention—and alternatives to retention.

C. IMPACT ON LOCAL CONTROL

Wise, A.E.

"Legislated Learning Revisited." ***Phi Delta Kappan*** **(Jan. 1988) pp. 328-333.**

Demonstrates how increasing use of standardized testing undermines local control over the public schools and increases state and national control over education. Also discusses other impacts such as narrowed curriculum, loss of teaching time, lower teacher morale, and loss of teacher authority.

5. AUTHENTIC EVALUATION AND REDUCED TESTING

The following materials are introductions to the concepts and essential elements of the practice of authentic, direct, performance-based evaluation. Detailed descriptions of possible ways to create systems of evaluation that are performance-based seem not yet to be generally available, excepting the North Carolina materials noted below, but hopefully will be available soon.

Archbald, D. and F. Newman

Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School **(Reston, VA: National Association of Secondary School Principals, 1988).**

Discusses "what is authentic academic achievement," how to assess it and educational programs, and implementing assessment

programs. Considers secondary school and college-level assessment alternatives currently in place. Appendix includes a critique of standardized tests.

Edelsky, C. & S. Harman

"One More Critique of Testing - With Two Differences." ***English Education*** **(Oct. 1988) pp. 157-171.**

Provides excellent summary of many problems with standardized tests, suggests appropriate assessment procedures to meet the different needs of parents, teachers and students; the public and elected officials; and researchers.

Educational Leadership

(Vol. 46, #7, April 1989) "Redirecting Assessment."

Includes 17 articles on developing authentic assessments, provides a wide range of materials useful for understanding alternatives theoretically and in practice. The best introduction to the range of "alternatives" being developed, most of which are performance-based.

Gardner, H.

"Assessment in Context: The Alternative to Standardized Testing." (Paper for the National Commission on Testing and Public Policy: Berkeley 1988).

Detailed discussion of process and product alternatives rooted in students' classroom work and recent scientific understandings. Gardner has recently published a number of other articles on assessment.

Haney, W.

"Making Testing More Educational." ***Educational Leadership*** **(October 1985) pp. 4-13.**

Describes the limited educational utility of standardized tests. Examines three efforts to make educational use of testing (Portland, OR; Orange County, FL; Pittsburgh, PA) and one school that uses no tests but relies on alternative evaluations (Prospect School, North Bennington, VT).

Johnston, P.

"Teachers as Evaluation Experts." ***The Reading Teacher*** **(April 1987) pp. 744-748.**

Discusses the fact that teachers can and do evaluate and urges that teachers need to be helped so that they can become evaluation experts.

North Carolina Department of Public Instruction

Grades 1 and 2 Assessment

N.C. has instituted developmentally appropriate assessment statewide in grades one and two, in Communication Skills and Mathematics. They have a packet of material available for $50.00, including a thick notebook, two videos, and material from staff development sessions. (116 W. Edenton St., Raleigh, N.C. 27603.)

Wiggins, G.

"A True Test: Toward More Authentic and Equitable Assessment." ***Phi Delta Kappan*** **(May 1989) pp. 703-713.**

Criticizes standardized, multiple-choice norm and criterion referenced tests as inadequate, inauthentic and harmful. Argues that authentic tests (as sometimes different from assessment in general) can be valuable for learning. Defines authentic tests as evidencing and providing a means to judge knowledge and gives examples. States that authentic tests enable interaction with the student and therefore enhance equity. Establishes criteria for authentic exams and for grading them. Links new forms of testing and assessment to re-structuring schools. (There are several articles on alternative assessments in this issue of *Kappan*.)

FairTest Publications

_____ ***FairTest Examiner***, published quarterly. . . $15.00/year (individuals), $25.00 (institutions)

_____ Back Issues of ***FairTest Examiner***, $4.00 each: Vol.___ No.___ (All 11 for $32.00)

_____ ***Standing Up to The SAT***, by John Weiss, Barbara Beckwith and Bob Schaeffer (192 pp.) . . . $6.95

_____ ***Fallout From the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools***, 3rd ed. by Noe Medina and Monty Neill (77 pp.). . . $8.95

_____ ***Sex Bias in College Admissions Tests: Why Women Lose Out***, by Phyllis Rosser and FairTest staff, 3rd ed. (60 pp.) $7.95

_____ ***The Reign of ETS***, by Allan Nairn and Ralph Nader (550 pp.). . . $30

_____ ***Beyond Standardized Tests: Admissions Alternatives That Work***, by Amy Allina with FairTest staff (18 pp.) $5.50

_____ ***None of the Above, Behind the Myth of Scholastic Aptitude***, by David Owen (327 pp.) . . . $5.95

Contributions to FairTest

I agree there's no better way to be part of the testing reform movement than supporting FairTest! Enclosed is my tax-deductible contribution!*

_____ **FairTest Associate** ($25.00--includes a free one-year subscription to the *FairTest Examiner*)

_____ **FairTest Sponsor** ($50.00--includes *Examiner* subscription and *Standing Up to the SAT*)

_____ **FairTest Sustainer** ($100.00 or more--includes above plus a 20% discount on publications)

_____ **Other**

Publications Total	_____
Contributions Total	_____
Check Total	_____

Name _____

Address _____

City _____ State _____ ZIP _____

Please make checks payable to FairTest and mail to: FairTest, 342 Broadway, Cambridge MA 02139

**Contributions to FairTest are fully tax deductible, except for the value of publications received.*